EIGENVALUE PROBLEMS FROM ELECTRONIC STRUCTURE THEORY

A thesis
submitted in fulfilment
of the requirements for the Degree
of
Master of Science in Mathematics
in the
University of Canterbury
by
Heather K. Jenkins

University of Canterbury
1998

Contents

List of Figures

List of Tables

Abstract

Eigenvalue problems from quantum chemistry are looked at. The topic is approached in such a way that a mathematician can understand not only the techniques used to solve these eigenproblems, but also their derivation which makes the meaning and usefulness of the results clearer. Various algorithms from both chemistry and mathematics are looked at.

A short review of eigenvalue problems from various areas of quantum chemistry is given and recent references are cited. Two particular eigenvalue problems are looked at in detail. Both come from looking at the electronic energy levels in molecules and are known as molecular orbital methods.

The first of these is in the self-consistent field procedure where Roothaan's equations are solved. The derivation of these equations is given along with the derivation of the Hartree-Fock equations which are needed to get Roothaan's equations. Level-shifting and direct inversion in the iterative space can both be used to improve the convergence of the procedure. Shepard's second-order SCF method for parallel implementation also improves convergence.

The second eigenvalue problem is in the configuration interaction method. The most common method used to solve this problem is Davidson's method. The Lanczos method is looked at and its relationship to Davidson's method is discussed. The convergence of Davidson's method and recent CI modifications are also explored.

Chapter 1

Introduction

The main goal of this thesis is to explain, to a mathematician, where some of the eigenvalue problems in computational chemistry come from.

The Chemistry Institute at ANL: In 1996 a *Large-Scale Matrix Diagonalization Methods in Chemistry Theory Institute* was held at Argonne National Laboratory. It brought together computational chemists and numerical analysts. The goal was to understand the needs of computational chemists in problems that use matrix diagonalization. A couple of publications that resulted from this workshop were useful as starting points for this thesis, [2] and [17]. Some important points from these references are now mentioned.

Davidson's method has continually dominated the problem of finding a few extremal eigenvalues for many computational chemistry problems since 1975 when it was first proposed. When a good preconditioner is available chemists tend to use this method. The fact that this method has been used for a long time, with only modest enhancements, shows how successful it is. Davidson's method is of particular interest in this thesis because it has at least been mentioned in various mathematical text books, like Parlett [26], Saad [32], and Trefethen and Bau [43]. However it has not been studied much by mathematicians which contrasts with the use made of the method by chemists.

The eigenvalue problems that were focused on come up in self-consistent field (SCF) theory, configuration interaction (CI), intramolecular vibrational relaxation (IVR) and scattering. It is interesting to note that methods for solving IVR and scattering problems, which require finding large numbers of eigenvalues and eigenvectors, have improved and changed in the last 20 years.

The chemists' matrices often have a couple of useful features. Chemists refer to their matrices as sparse since only a few percent of the elements are non-zero. Numerical analysts will refer to a matrix as sparse if the number of non-zeros is not proportional to the matrix size. So the chemists matrices have few non-zeros but are not usually sparse in the traditional matrix theory sense. Another important property of their matrices is that they are often diagonally dominant. Both of these features can greatly simplify things.

Some of the techniques used by chemists are not guaranteed to produce the right answer, but in practice the correct solutions seem to be found. This is true for Davidson's method as the initial guess is sometimes not close enough to the final solution to guarantee convergence.

At the workshop numerical analysts expressed interest in creating sample test cases that represent real chemistry problems. Another area of interest is how to incorporate the insights on the nature of the problems into general eigensolvers. A general package needs to include preconditioners that are as good as the physical knowledge currently used for specific problems. The institute showed that chemists, numerical analysts and computer scientists will have to work together if the chemistry problems are going to get solved.

Eigenvalue Problems in Quantum Chemistry: We now give an overview of some of the eigenvalue problems in chemistry.

The biggest single area of chemistry where eigenvalue problems have to be solved is in determining the electronic energy levels of molecules. This is done using molecular orbital methods. The orbital concept is an approximation and it is used for the qualitative discussion of the chemistry of atoms and molecules. Molecular orbital theory gives us a way to think about molecular electronic structure. The theory has been developing rapidly since the 1950s when computers became available. The matrices are almost always very large and therefore specialized methods are used to solve the eigenvalue problems. Davidson's method is commonly used in this area and it is included in many computer packages that are used by chemists. SCF and CI methods are a part of this area. It is these two eigenvalue problems that are focused on in this thesis. The theory behind the CI method builds on that of the SCF procedure, so SCF is looked at before CI. The actual computational methods that are talked about for SCF and CI are not connected in this way. The chemistry and physics involved in SCF theory is relatively easy to understand and for this reason we look at it in detail. Also by doing this a lot of notation can be defined, and basic general concepts used by chemists can be explained. It is important to note that the theory needed for the electronic structure methods that are not discussed here is more complicated but it does in many ways follow on from what we do look at here. The bibliography includes references for these harder methods.

SCF and CI calculations are both done using an approach based on the Rayleigh-Ritz variation method, which is looked at in subsection 2.1.4. The approximation they use is a linear expansion of the wave function in a finite dimensional function space and the numerical solution involves a symmetric matrix eigenvalue problem. In the SCF method the symmetric matrices range from order hundreds to thousands. These matrices often include large clusters of eigenvalues that can be as much as 25 % of the spectrum. With CI methods the matrix size can be between 10^4 and 10^9 where only a few extremal eigenpairs are needed. Working with such large matrices has led to the development of specialized methods. Parallel computers have meant that problems of size 10^9 have been solved.

Eigenvalue problems are solved a lot in determining the vibrational energy levels of poly-atomic molecules. The same ideas as are used in molecular orbitals get used here but are applied to vibration states rather than electronic states. The matrix that is diagonalized has different characteristics so usually different methods get used. It can come from a time-independent Schrödinger equation with a vibrational Hamiltonian. In this area another thing of interest is intramolecular energy flow and this involves solving the time-dependent Schrödinger equation. These problems are described in [49]. IVR fits into this general area.

Another area of study is collisions between molecules. Often the problem can be reduced to something that looks like a vibrational eigenvalue problem and therefore the methods used are similar to the vibrational ones. Recent references in this area are [48] and [50]. Eigenvalue methods are also important in chemical kinetics and this area is discussed in [9].

What Is Coming Up: In the next chapter the Hartree-Fock equations are looked at. These equations come from taking the time-independent Schrödinger equation for molecules and making a simple approximation. The chemistry that is needed to understand the Hartree-Fock equations is given. As they stand these equations can only be solved for very simple systems and further approximations need to be made.

In chapter 3 more approximations are made and we get the matrix equations known as the Roothaan equations. The SCF procedure, which is used to solve these equations, is described. The material in this chapter shows what is involved, in terms of approximations and computations, in going from some unsolvable equations to matrix equations that can be solved on a computer. In a sense we are going from quantum chemistry to computational chemistry. The ideas in this chapter are the key to understanding what goes into solving these sorts of chemistry problems.

Configuration interaction is the main topic of chapter 4. Here we look at getting better results than those of the Hartree-Fock approximation. The idea of the CI method is to use a linear combination of wave functions that are the result of doing a calculation like that in chapter 3. A calculation is done to choose the best expansion coefficients.

In chapter 5 ways of improving the SCF method are discussed. We look at procedures that are currently used to improve the convergence of the SCF procedure as it is given in chapter 3. A recent SCF method that does not involve forming the Roothaan equations is looked at with regard to parallel implementation.

Davidson's method is reached in chapter 6. Some mathematical background is needed for Davidson's method and this includes the Lanczos method. Davidson's method has been generalized and the generalized version includes the Lanczos method. We look at the CI method in more detail. In particular we consider the CI eigenvalue problem and chemists tend to use Davidson's method to solve this problem. The convergence of the method is looked at with the CI eigenvalue problem in mind. Some recent modifications of Davidson's method that are used for CI are discussed.

Finally in chapter 7 a short summary is given.

Table 1.1 lists the algorithms that will be looked at throughout the chapters. The line divides the chemistry procedures from the mathematical procedures.

procedure	*page*
self-consistent field (SCF)	43
level-shifting	73
direct inversion in the iterative subspace (DIIS)	79
parallel direct SCF	89
conventional CI	95
Rayleigh-Ritz	101
Lanczos	106
Generalized Davidson method	112
Davidson's (original) method	114

Table 1.1: The schemes that are outlined.

A background in undergraduate varsity mathematics as well as high school chemistry and physics is needed in order to read this thesis.

Chapter 2

The Hartree-Fock Equations

In this chapter we are going to be looking at the eigenvalue problem of the Hartree-Fock equations. The Hartree-Fock equations are a part of electronic structure theory and this is one of the areas of chemistry where computational methods provide information that is complementary to experimental results. The type of calculation we are going to look at is an example of an *ab initio* (which means from the beginning) calculation and it is so called because it is done without using experimental data. Generally speaking, *ab initio* methods are potentially capable of reproducing experimental results without using empirical parameters. They can also give insights into a problem that cannot be obtained from an experiment. The lack of experimental knowledge also means that the ideas can be more easily understood by someone who is not a chemist. Calculations that use experimental data automatically require a knowledge of more chemistry.

The Hartree-Fock approximation is the molecular orbital approximation, that is the approximation that electrons in molecules occupy orbitals. It is used as a starting point for more accurate approximations that include the effects of electron correlation. We will look at what electron correlation is in subsection 2.1.8. The methods that include correlation effects are discussed in Szabo and Ostlund [42], and Hirst [16] for example. We will look at some of these methods in later chapters.

In this chapter we derive the Hartree-Fock equations. Before doing this some background material is covered. We begin by looking at the quantum theory behind the electronic structure of atoms and molecules. The sort of notation chemists use is defined, and the Schrödinger equation that we are interested in solving is given.

In the next chapter we look at how the Hartree-Fock equations can be put into a solvable form for a particular case. The Roothaan equations allow us to calculate Hartree-Fock solutions for the ground state of closed-shell molecules. We will look at what the terms ground state and closed-shell mean later in subsection 3.1.1. The self-consistent field procedure (SCF) described is the most basic method used to solve the Roothaan equations.

The book by Szabo and Ostlund [42] was used extensively in preparing this chapter and the one that follows.

2.1 Background Material

Before we can look at where quantum chemists get any of their eigenvalue problems from we need to review some basic ideas about the electronic structure of atoms. The Pauli exclusion principle and electron correlation also need to be looked at. Bracket notation and Slater determinants are defined. For any additional background chemistry the books [45] and [20] are useful.

2.1.1 Basic Quantum Chemistry

We know that the electrons in atoms can occupy certain discrete energy levels [45]. Consequently electrons absorb or emit energy in discrete amounts as they move from one energy level to another. It is more effective to treat electrons in atoms as waves. Quantum mechanics describes the behaviour of very small particles and is based on the wave properties of matter. Quantization of energy is a consequence of these properties. The Heisenberg Uncertainty principle is important. It states that it is impossible to determine accurately the momentum and position of an electron, or any other very small particle, simultaneously. So we can only talk about the probability of finding an electron within a specified region of space.

Each solution of the Schrödinger wave equation, which is described in subsection 2.1.3, gives a possible energy state for the electrons in the atom. Also the allowed energy states of atoms or molecules can be described by sets of numbers called **quantum numbers**. The Schrödinger equation is a second-order differential equation and it has only been solved exactly for hydrogen. Solutions of it also tell us about the shapes and orientations of the statistical probability distributions of the electrons. An **atomic orbital** is a region of space in which the probability of finding an electron is high. The atomic orbitals are deduced from the solutions of the Schrödinger equation and are directly related to the quantum numbers. We can use the quantum numbers to describe the electronic arrangements in all atoms, their so called electronic configurations. The discussion of quantum numbers is deferred until subsection 3.3.2, which is where it is needed. Also that is a good place for a change of pace.

2.1.2 Bracket Notation

This notation was introduced by Dirac. Consider N basis vectors $|i\rangle, i = 1, 2, \ldots, N$ which are **ket vectors** or **kets**. As long as the basis is complete any ket $|a\rangle$ can be expressed as

$$|a\rangle = \sum_{i=1}^{N} |i\rangle a_i \tag{2.1}$$

We can represent $|a\rangle$ by the vector

$$\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_N \end{pmatrix}$$

The adjoint of $\mathbf{a}$ is $\mathbf{a}^H = (a_1^* \, a_2^* \, \ldots \, a_N^*)$, where a_i^* is the complex conjugate of a_i. A **bra vector** $\langle a|$ has a matrix representation $\mathbf{a}^H$. The **scalar product** between a bra $\langle a|$ and a ket $|b\rangle$ is

$$\langle a|b\rangle \;\equiv\; \langle a||b\rangle \;=\; \mathbf{a}^H\mathbf{b} \;=\; \sum_{i=1}^{N} a_i^* \, b_i \tag{2.2}$$

Note that later in equation (2.27), when we are talking about double integrals, $\langle ij|kl\rangle$ and $\langle ij||kl\rangle$ will not be the same. This notation is also used when the basis set is infinite. We can write

$$\langle a| = \sum_i a_i^* \langle i|$$

so that

$$\langle a|b\rangle = \sum_{ij} a_i^* \langle i|j\rangle b_j$$

and hence

$$\langle i|j\rangle = \delta_{ij}$$

In other words we must have *orthonormality*. The components of the vectors are given by

$$a_j = \sum_i \delta_{ji} a_i = \sum_i \langle j|i\rangle a_i = \langle j|a\rangle$$

and

$$a_j^* = \sum_i a_i^* \delta_{ij} = \sum_i a_i^* \langle i|j\rangle = \langle a|j\rangle$$

as we would expect. So we can express equation (2.1) as

$$|a\rangle = \sum_i |i\rangle a_i = \sum_i |i\rangle\langle i||a\rangle$$

and so we can put

$$1 = \sum_i |i\rangle\langle i|$$

and we think of this in terms of the way it acts on a ket vector. More generally an **operator** $\mathcal{O}$ converts a ket vector into another ket ie.

$$\mathcal{O}|a\rangle = |b\rangle \tag{2.3}$$

We can represent $\mathcal{O}$ by the matrix O if we know how it transforms the basis $\{|i\rangle\}$

$$\mathcal{O}|i\rangle = \sum_j O_{ji}|j\rangle$$

From this we get

$$O_{ki} = \sum_j \delta_{kj} O_{ji} = \sum_j \langle k|j\rangle O_{ji} = \langle k|\mathcal{O}|i\rangle$$

The **adjoint** $\mathcal{O}^\dagger$ is defined by

$$\langle a|\mathcal{O}^\dagger = \langle b| \tag{2.4}$$

and it is represented by O^H.

Bracket notation is also used for functions in the following way,

$$a(x) \equiv |a\rangle \qquad\qquad a^*(x) \equiv \langle a|$$

and

$$\phi_i(x) \equiv |i\rangle \qquad\qquad \phi_i^*(x) \equiv \langle i|$$

The scalar product of two functions is

$$\langle a|b\rangle \equiv \int a^*(x)b(x)dx \tag{2.5}$$

Also

$$\langle a|\mathcal{O}|b\rangle \equiv \int a^*(x)\mathcal{O}b(x)dx \tag{2.6}$$

The integrals are over the region of interest.

2.1.3 The Time-Independent Schrödinger Equation

The **non-relativistic time-independent Schrödinger equation** is the eigenvalue equation

$$\mathcal{H}|\Phi\rangle = \mathcal{E}|\Phi\rangle, \tag{2.7}$$

where $\mathcal{H}$ is the **Hamiltonian** which is a Hermitian energy operator, $|\Phi\rangle$ is the **wave function**, and $\mathcal{E}$ is the **energy**.

In quantum mechanics wave functions are used to describe the state of a system. Here we are describing electrons in an atom or a molecule. The existence of such a function is one of the postulates of quantum mechanics [20, page 9]. It describes the system such that $|\Phi|^2$ is a probability density function for the position of an electron.

The equation can only be solved exactly in simple cases. There is an infinite set of exact solutions to the Schrödinger equation

$$\mathcal{H}|\Phi_\alpha\rangle = \mathcal{E}_\alpha|\Phi_\alpha\rangle \qquad \alpha = 0, 1, \ldots \tag{2.8}$$

where $\mathcal{E}_0 \leq \mathcal{E}_1 \leq \ldots \leq \mathcal{E}_\alpha \leq \ldots$. These eigenvalues give the only possible values of the energy of the system. For simplicity we have assumed that the set of the eigenvalues is discrete. Since $\mathcal{H}$ is Hermitian the eigenvalues $\{\mathcal{E}_\alpha\}$ are real and the eigenfunctions are orthonormal

$$\langle\Phi_\alpha|\Phi_\beta\rangle = \delta_{\alpha\beta}$$

Therefore by multiplying by $\langle\Phi_\beta|$ on the left hand side (2.8) becomes

$$\langle\Phi_\beta|\mathcal{H}|\Phi_\alpha\rangle = \mathcal{E}_\alpha\delta_{\alpha\beta}$$

which means that the eigenvalues are given by

$$\langle \Phi_\alpha | \mathcal{H} | \Phi_\alpha \rangle = \mathcal{E}_\alpha$$

We also assume that the eigenfunctions of $\mathcal{H}$ form a complete set, and therefore any function $|\widetilde{\Phi}\rangle$ that satisfies the same boundary conditions as the set $\{|\Phi_\alpha\rangle\}$ can be expressed as a linear combination of the $|\Phi_\alpha\rangle$'s ie.

$$|\widetilde{\Phi}\rangle = \sum_\alpha |\Phi_\alpha\rangle c_\alpha = \sum_\alpha |\Phi_\alpha\rangle\langle \Phi_\alpha|\widetilde{\Phi}\rangle$$

and

$$\langle\widetilde{\Phi}| = \sum_\alpha c_\alpha^* \langle \Phi_\alpha| = \sum_\alpha \langle\widetilde{\Phi}|\Phi_\alpha\rangle\langle\Phi_\alpha|$$

We usually want the wave function to vanish at infinity.

The **expectation value** of the Hamiltonian, for a wave function $|\widetilde{\Phi}\rangle$, is

$$E_0 \equiv \langle\widetilde{\Phi}|\mathcal{H}|\widetilde{\Phi}\rangle \tag{2.9}$$

2.1.4 The Variation Principle

Theorem: *If a normalized wave function $|\widetilde{\Phi}\rangle$ satisfies the appropriate boundary conditions, then the expectation value of the Hamiltonian is an upper bound to the exact ground state energy. In other words if $\langle\widetilde{\Phi}|\widetilde{\Phi}\rangle = 1$ then*

$$\langle\widetilde{\Phi}|\mathcal{H}|\widetilde{\Phi}\rangle \geq \mathcal{E}_0$$

We have equality only when $|\widetilde{\Phi}\rangle = |\Phi_0\rangle$.

This theorem is called the **(Rayleigh-Ritz) variation principle**. The ground-state gives the minimum expectation value.

The proof is easy to follow and illustrates bracket notation. Consider

$$\begin{aligned}\langle\widetilde{\Phi}|\widetilde{\Phi}\rangle = 1 = \sum_{\alpha\beta}\langle\widetilde{\Phi}|\Phi_\alpha\rangle\langle\Phi_\alpha|\Phi_\beta\rangle\langle\Phi_\beta|\widetilde{\Phi}\rangle &= \sum_{\alpha\beta}\langle\widetilde{\Phi}|\Phi_\alpha\rangle\delta_{\alpha\beta}\langle\Phi_\beta|\widetilde{\Phi}\rangle \\ &= \sum_\alpha \langle\widetilde{\Phi}|\Phi_\alpha\rangle\langle\Phi_\alpha|\widetilde{\Phi}\rangle = \sum_\alpha |\langle\Phi_\alpha|\widetilde{\Phi}\rangle|^2\end{aligned}$$

and recall that $\mathcal{E}_\alpha \geq \mathcal{E}_0$ for all α. Thus

$$\begin{aligned}\langle\widetilde{\Phi}|\mathcal{H}|\widetilde{\Phi}\rangle &= \sum_{\alpha\beta}\langle\widetilde{\Phi}|\Phi_\alpha\rangle\langle\Phi_\alpha|\mathcal{H}|\Phi_\beta\rangle\langle\Phi_\beta|\widetilde{\Phi}\rangle = \sum_{\alpha\beta}\langle\widetilde{\Phi}|\Phi_\alpha\rangle\mathcal{E}_\beta\delta_{\alpha\beta}\langle\Phi_\beta|\widetilde{\Phi}\rangle \\ &= \sum_\alpha \mathcal{E}_\alpha|\langle\Phi_\alpha|\widetilde{\Phi}\rangle|^2 \geq \sum_\alpha \mathcal{E}_0|\langle\Phi_\alpha|\widetilde{\Phi}\rangle|^2 = \mathcal{E}_0\sum_\alpha|\langle\Phi_\alpha|\widetilde{\Phi}\rangle|^2 = \mathcal{E}_0 \qquad \blacksquare\end{aligned}$$

Suppose we are approximating the exact ground state wave function $|\Phi_0\rangle$ by $|\widetilde{\Phi}\rangle$. It can be shown [35, page 193] that the error in the expectation value is second order with respect to the error in the wave function. This means if $\varepsilon|\Delta\rangle = |\widetilde{\Phi}\rangle - |\Phi_0\rangle$, where $\langle\Delta|\Delta\rangle = 1$, then

$$E_0 = \langle\widetilde{\Phi}|\mathcal{H}|\widetilde{\Phi}\rangle = \mathcal{E}_0 + \varepsilon^2\langle\Delta|(\mathcal{H} - \mathcal{E}_0)|\Delta\rangle \tag{2.10}$$

This idea will be important later on when we look at modifying Davidson's method.

2.1.5 The Electronic Hamiltonian

We are interested in finding approximate solutions of equation (2.7) for systems of nuclei and electrons, which we describe by position vectors $\mathbf{R}_A$ and $\mathbf{r}_i$ respectively. The Hamiltonian is given in atomic units and this means that equation has been scaled so that it is dimensionless. For N electrons and M nuclei it is

$$\mathcal{H} = -\sum_{i=1}^{N} \tfrac{1}{2}\nabla_i^2 - \sum_{A=1}^{M} \frac{1}{2M_A}\nabla_A^2 - \sum_{i=1}^{N}\sum_{A=1}^{M} \frac{Z_A}{r_{iA}} + \sum_{i=1}^{N}\sum_{j>i}^{N} \frac{1}{r_{ij}} + \sum_{A=1}^{M}\sum_{B>A}^{M} \frac{Z_A Z_B}{R_{AB}} \tag{2.11}$$

where M_A is the mass of the nucleus A relative to the mass of an electron, Z_A is the atomic number, ∇_i^2, ∇_A^2 are Laplacians and r_{iA} is the distance between the ith electron and the Ath nucleus. Part of the coordinate system is shown in figure 2.1.

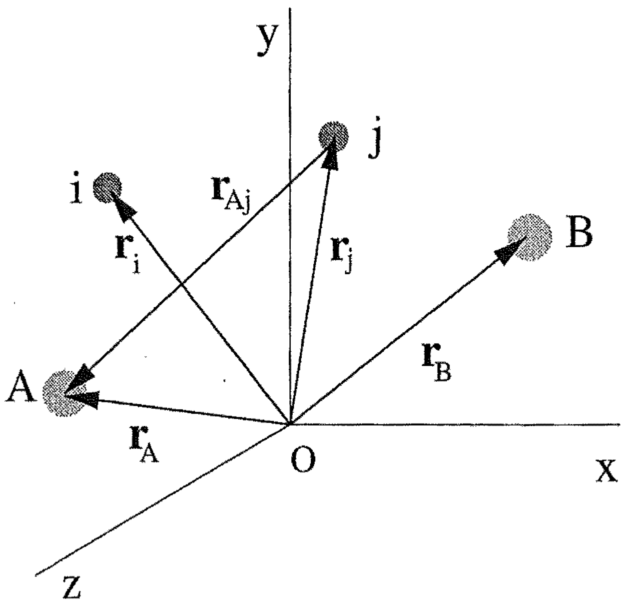


Figure 2.1: The coordinate system for electrons i and j, and nuclei A and B.

Let us look at what the terms in the equation mean. The first two terms are the kinetic energy operators. The coulomb attraction between electrons and nuclei is represented by the third term. Finally the last two terms are for the repulsion between electrons and nuclei respectively. So in this case the Schrödinger equation is a linear partial differential equation.

We can simplify the equation above by making the approximation that the nuclei are fixed. This is known as the **Born-Oppenheimer approximation** and it is valid because the nuclei are much heavier than the electrons. This means we can ignore the second term of (2.11) and the last term is constant. What we have left of (2.11) is called the **electronic Hamiltonian** and it represents the motion of N electrons in a field of M point charges.

$$\mathcal{H}_{\text{elec}} = -\sum_{i=1}^{N} \tfrac{1}{2}\nabla_i^2 - \sum_{i=1}^{N}\sum_{A=1}^{M} \frac{Z_A}{r_{iA}} + \sum_{i=1}^{N}\sum_{j>i}^{N} \frac{1}{r_{ij}} \tag{2.12}$$

We have ignored the constant term because it does not change the eigenfunctions and it adds a constant to the eigenvalues. A solution Φ_{elec} of

$$\mathcal{H}_{\text{elec}}\Phi_{\text{elec}} = \mathcal{E}_{\text{elec}}\Phi_{\text{elec}} \tag{2.13}$$

is an **electronic wave function**, and it depends *explicitly* on the electronic coordinates and *parametrically* on the nuclear coordinates.

If we could solve equation (2.7) for a molecule we would be able to get all the information about the molecule consistent with the postulates of quantum mechanics. However we can only do this for simple systems and we end up having to make approximations. The first of these is in replacing (2.7) by (2.13).

To completely describe an electron we need to specify its spin. Let $\alpha(\omega)$ and $\beta(\omega)$ be the **spin functions** and they correspond to spin up and spin down respectively. In subsection 3.3.2 we look at spin and the spin quantum number. The variable ω is called the **spin variable** and it will not be specified. We need to have the two spin functions being complete and orthonormal, so that

$$\langle\alpha|\alpha\rangle = 1 = \langle\beta|\beta\rangle$$

and

$$\langle\alpha|\beta\rangle = 0 = \langle\beta|\alpha\rangle$$

Now an electron is described by the ordered pair $\mathbf{x} = \{\mathbf{r}, \omega\}$ and the wave function for an N electron system depends on $\mathbf{x}_1, \ldots, \mathbf{x}_N$. The Hamiltonian does not depend on spin so this does not change anything at this stage.

2.1.6 The Pauli Exclusion Principle

Due to the uncertainty principle identical particles like electrons are indistinguishable when they are in the same system, and therefore we cannot tell which electron is in which orbital. If two electrons are well separated so that their wave functions do not overlap then we can distinguish between them.

Suppose P_{ij} is defined by

$$P_{ij}\, f(\mathbf{x}_1, \ldots, \mathbf{x}_i, \ldots, \mathbf{x}_j, \ldots, \mathbf{x}_N) = f(\mathbf{x}_1, \ldots, \mathbf{x}_j, \ldots, \mathbf{x}_i, \ldots, \mathbf{x}_N) \tag{2.14}$$

Then 1 is an eigenvalue of P_{ij}^2 so that P_{ij} has eigenvalues 1 and -1. So the eigenfunctions of P_{ij} are symmetric and antisymmetric respectively. In the same way as we can express any matrix as the sum of a symmetric matrix and a skew-symmetric matrix, we can write any function f as

$$\begin{aligned} f(\ldots, \mathbf{x}_i, \ldots, \mathbf{x}_j, \ldots) &= \tfrac{1}{2}\left[\, f(\ldots, \mathbf{x}_i, \ldots, \mathbf{x}_j, \ldots) + f(\ldots, \mathbf{x}_j, \ldots, \mathbf{x}_i, \ldots)\,\right] \\ &+ \tfrac{1}{2}\left[\, f(\ldots, \mathbf{x}_i, \ldots, \mathbf{x}_j, \ldots) - f(\ldots, \mathbf{x}_j, \ldots, \mathbf{x}_i, \ldots)\,\right] \end{aligned}$$

Consequently the eigenfunctions form a complete set.

Since electrons are indistinguishable the way they are labelled cannot affect the state of a system, so $\psi(\ldots, \mathbf{x}_i, \ldots, \mathbf{x}_j, \ldots)$ and $\psi(\ldots, \mathbf{x}_j, \ldots, \mathbf{x}_i, \ldots)$ correspond to the same state. Therefore they have a constant ratio and differ by a constant factor of c, and so

$$P_{ij}\, \psi(\ldots, \mathbf{x}_i, \ldots, \mathbf{x}_j, \ldots) \;=\; \psi(\ldots, \mathbf{x}_j, \ldots, \mathbf{x}_i, \ldots) \;=\; c\, \psi(\ldots, \mathbf{x}_i, \ldots, \mathbf{x}_j, \ldots)$$

So ψ is an eigenfunction and therefore $c = \pm 1$. Hence a wave function has to be either symmetric or antisymmetric.

Experimental evidence has lead to the postulate that *the wave function of a system of electrons must be antisymmetric with respect to the interchange of any two electrons.* This is the **Pauli exclusion principle** or the **antisymmetry principle**.

So we need the wave function to satisfy

$$\psi(\mathbf{x}_1, \ldots, \mathbf{x}_i, \ldots, \mathbf{x}_j, \ldots, \mathbf{x}_N) = -\psi(\mathbf{x}_1, \ldots, \mathbf{x}_j, \ldots, \mathbf{x}_i, \ldots, \mathbf{x}_N)$$

If $\mathbf{x}_1 = \mathbf{x}_2$ then

$$\psi(\mathbf{x}_1, \mathbf{x}_1, \mathbf{x}_3, \ldots, \mathbf{x}_N) = -\psi(\mathbf{x}_1, \mathbf{x}_1, \mathbf{x}_3, \ldots, \mathbf{x}_N)$$

and so $\psi = 0$. Hence electrons with the same spin have zero probability of being found at the same point in space. Since ψ is continuous this means that the probability of finding electrons with the same spin close together is small. This gets referred to as **Pauli repulsion** but it is not a real physical force.

The Pauli exclusion principle was originally stated as; *no two electrons can occupy the same spin-orbital.* The above is more general than this. This means that no two electrons in an atom can have the same set of four quantum numbers. Consequently each orbital ψ_i can contain a maximum of two electrons, one with spin function α and the other with spin function β.

So far we have that the exact wave function has to satisfy the Schrödinger equation and must be antisymmetric. By using Slater determinants it is easy to satisfy antisymmetry and we use them to describe many-electron wave functions.

2.1.7 Slater Determinants

By definition an **orbital** is a wave function for an electron. **Molecular orbitals** are the wave functions of electrons in a molecule. The **spatial orbital** $\psi(\mathbf{r})$ describes the spatial distribution of an electron such that $|\psi(\mathbf{r})|^2 d\mathbf{r}$ is the probability of finding an electron in the small volume $d\mathbf{r}$ at $\mathbf{r}$. This is the condition used to normalize ψ. Spatial molecular orbitals form an orthonormal set. A **spin orbital** $\chi(\mathbf{x})$ is the wave function that describes spatial distribution and spin so that

$$\chi(\mathbf{x}) \equiv \begin{cases} \psi(\mathbf{r})\alpha(\omega) \\ \quad\text{or} \\ \psi(\mathbf{r})\beta(\omega) \end{cases} \tag{2.15}$$

When the spatial orbitals are orthonormal so are the spin orbitals.

Before looking at Slater determinants we will look at Hartree products. If we neglect electron-electron repulsions the electronic Hamiltonian has the form

$$\widetilde{\mathcal{H}}_{\text{elec}} = \sum_{i=1}^{N} h(i) \tag{2.16}$$

where $h(i)$ is the operator that describes the kinetic and potential energy of the ith electron. It is defined by

$$h(i) \equiv -\tfrac{1}{2}\nabla_i^2 - \sum_A \frac{Z_A}{|\mathbf{r}_i - \mathbf{R}_A|} \tag{2.17}$$

The set of spin orbitals $\{\chi_j\}$ are the eigenfunctions for $h(i)$ and we have

$$h(i)\chi_j(\mathbf{x}_i) = \mathcal{E}_j\chi_j(\mathbf{x}_i) \tag{2.18}$$

The many-electron wave function Ψ^{HP} is a **Hartree product** and is given by

$$\Psi^{\text{HP}}(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N,) \equiv \chi_k(\mathbf{x}_1)\chi_l(\mathbf{x}_2)\ldots\chi_m(\mathbf{x}_N) \tag{2.19}$$

It is an eigenfunction of $\widetilde{\mathcal{H}}_{\text{elec}}$

$$\widetilde{\mathcal{H}}_{\text{elec}}\,\Psi^{\text{HP}} = E\,\Psi^{\text{HP}}$$

with

$$E = \mathcal{E}_k + \mathcal{E}_l + \cdots + \mathcal{E}_m$$

The definition of the Hartree product makes sense if we consider the probability interpretation of χ and assume that the spin orbitals are in some sense independent.

The Hartree product has electron-one occupying spin orbital χ_k, electron-two in χ_l, $\ldots$, but the electrons are in fact indistinguishable according to the uncertainty principal. Also the Hartree product is not antisymmetric with respect to the interchange of coordinates. This motivates the definition of a **(single) Slater determinant** which is

$$\Psi(\mathbf{x}_1, \ldots, \mathbf{x}_N) \equiv (N!)^{-\frac{1}{2}} \begin{vmatrix} \chi_k(\mathbf{x}_1) & \chi_l(\mathbf{x}_1) & \ldots & \chi_m(\mathbf{x}_1) \\ \chi_k(\mathbf{x}_2) & \chi_l(\mathbf{x}_2) & \ldots & \chi_m(\mathbf{x}_2) \\ \vdots & \vdots & & \vdots \\ \chi_k(\mathbf{x}_N) & \chi_l(\mathbf{x}_N) & \ldots & \chi_m(\mathbf{x}_N) \end{vmatrix} \tag{2.20}$$

We have N electrons occupying N spin orbitals without specifying which electron is in which orbital. If we interchange the coordinates of two electrons two rows of the determinant are interchanged and therefore the antisymmetry principal is satisfied. If two electrons occupy the same orbital the determinant is zero which is consistent with the original Pauli exclusion principal. Recall that we want $|\Psi|^2$ to be the probability density function for the position of an electron. The normalization constant $(N!)^{-\frac{1}{2}}$ is necessary to make this true. The Slater determinant is the simplest antisymmetric wave function which can be used to describe the ground states of an N–electron system.

We use the notation

$$|\Psi\rangle = \Psi(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N) = |\chi_k(\mathbf{x}_1)\chi_l(\mathbf{x}_2)\cdots\chi_m(\mathbf{x}_N)\rangle = |\chi_k\chi_l\cdots\chi_m\rangle$$

so that antisymmetry can be expressed as

$$|\cdots\chi_i\cdots\chi_j\cdots\rangle = -|\cdots\chi_j\cdots\chi_i\cdots\rangle$$

A Hartree product is an independent-electron wave function because the probability of finding electron-one in $d\mathbf{x}_1$ at $\mathbf{x}_1$, electron-two in $d\mathbf{x}_2$ at $\mathbf{x}_2$ etc. , is equal to the product of the probabilities that electron-one is in $d\mathbf{x}_1$, electron-two is in $d\mathbf{x}_2$, etc. We get a Slater determinant by antisymmetrizing a Hartree product, and as we shall we it incorporates **exchange correlation**, which means that the motion of electrons with parallel spins is correlated. The motion of electrons with opposite spins is not. Let us now look at electron correlation in more detail.

2.1.8 Electron Correlation

Energies that are calculated using the Hartree-Fock method are typically in error by approximately 1 % [20, page 265]. This error is not acceptable when chemists are calculating things like bond energies, which involves taking the difference between quantities.

A Hartree-Fock SCF wave function averages the interactions between electrons, but it is necessary to consider the instantaneous interaction between electrons. Electrons tend to avoid each other because of the repulsion between them. It is energetically more favourable for electrons to be separated. There is a **Coulomb hole** surrounding each electron in an atom. This is a volume where the probability of finding another electron is small. Therefore the *motion of electrons is correlated* with each other. To improve the Hartree-Fock approximation we need to introduce the instantaneous electron correlation into the wave function.

A Hartree-Fock wave function satisfies the antisymmetry principle and therefore vanishes when two electrons with the same spin have the same spatial coordinates. Also because of continuity there is little probability of finding electrons of the same spin in the same region of space. Hence it includes some correlation of the motions of electrons with the same spin. This is exchange correlation. We refer to a **Fermi hole** around each electron and this is a region in space where the probability of finding another electron with the same spin is small.

When the wave function is constrained to have doubly occupied molecular orbitals inter-electronic repulsion cannot be taken into account fully. The use of doubly occupied orbitals in the Hartree-Fock method has another serious disadvantage for molecules. *Dissociation is not described correctly.* A molecular wave function in which the orbitals are always doubly occupied cannot dissociate into two fragments with each containing a singly occupied orbital. In terms of what we are discussing here, this means if we perform a calculation with the nuclear coordinates such that the nuclei are well separated, then the result will be of little value.

The **correlation energy** E_{corr} is the difference between the energy of the Hartree-Fock wave function E_0 and the true non-relativistic energy $\mathcal{E}_0$.

$$E_{\text{corr}} \equiv \mathcal{E}_0 - E_0 \tag{2.21}$$

It is always negative.

There are two main ways of allowing for instantaneous electron correlation. The first is to introduce inter-electronic distances r_{ij}, but this is only practical for systems with a small number of electrons. The second is **configuration interaction** (CI) which is also known as **configuration mixing** (CM). This is discussed in section 4.2.

Electron correlation effects are divided into two categories. **Dynamical correlation** is the correlation between the motion of electrons arising because of the coulomb interaction between electrons. **Non-dynamical correlation** refers to other problems with the wave function, like the inability to describe dissociation correctly.

The Hartree-Fock approximation will get less and less accurate as the atoms are separated. Because the correlation energy also includes nondynamical effects the Hartree-Fock approximation often leads to the result that the correlation energy increases as the electrons move apart as the atoms separate. This is not intuitive.

2.2 The Hartree-Fock Equations

Using a technique called functional variation we will derive the **Hartree-Fock** equations in their general spin orbital form. That is, we get the equation

$$f|\chi_a\rangle = \varepsilon_a|\chi_a\rangle \tag{2.22}$$

by minimizing the energy expression for a single Slater determinant. Here f is the Fock operator which is defined below in equation (2.33). The Hartree-Fock approximation replaces a complicated many-electron problem by a one-electron problem. It is due to Hartree [15] and Fock [8]. Electron-electron repulsion is treated in an average way by f. Also f depends on its eigenfunctions so that this equation is non-linear and hence must be solved using iterative methods. The spin orbitals $\{\chi_a\}$ that satisfy this equation give the single determinant $|\Psi_0\rangle = |\chi_1\chi_2\ldots\chi_N\rangle$, which is the best approximation to the ground state (hence the zero subscript) of the N–electron system described by the electronic Hamiltonian $\mathcal{H}_{\text{elec}}$ of equation (2.12).

There are two important associated theorems. The first is Koopmans' theorem which is an interpretation of the Hartree-Fock orbital energies (or eigenvalues) as ionization potentials and electron affinities. Secondly we have Brillouin's theorem which states that the matrix element between a Hartree-Fock single determinant and determinants that differ by a single excitation is zero.

Atoms have spherical symmetry and it is possible to solve the Hartree-Fock equations numerically to give the atomic orbitals $\{\psi_i\}$. However this is not possible for molecules which have lower symmetry. So we need to introduce a basis and get the Roothaan equations. This is done in the next chapter.

2.2.1 Minimizing the Electronic Energy

By the variation principle the best spin orbitals minimize the electronic energy, and their expectation value is

$$\begin{aligned} E_0 = \langle\Psi_0|\mathcal{H}_{\text{elec}}|\Psi_0\rangle &= \sum_a \langle a|h|a\rangle + \tfrac{1}{2}\sum_{ab}\langle ab||ab\rangle \\ &= \sum_a [a|h|a] + \tfrac{1}{2}\sum_{ab}[aa|bb] - [ab|ba] \end{aligned}$$

Let us define this new notation which is used for *spin orbitals*. Firstly

$$[i|j] \equiv \langle i|j\rangle \tag{2.23}$$

$$[i|h|j] \equiv \langle i|h|j\rangle \tag{2.24}$$

where the right hand sides of the above equations have already been defined. Recall that h is the operator that describes the kinetic and potential energy of an electron and is given by equation (2.17). Now the notation deviates

$$\langle ij|kl\rangle = \langle \chi_i\chi_j|\chi_k\chi_l\rangle \equiv \int \chi_i^*(\mathbf{x}_1)\chi_j^*(\mathbf{x}_2)r_{12}^{-1}\chi_k(\mathbf{x}_1)\chi_l(\mathbf{x}_2)d\mathbf{x}_1 d\mathbf{x}_2 \tag{2.25}$$

$$[ij|kl] = [\chi_i\chi_j|\chi_k\chi_l] \equiv \int \chi_i^*(\mathbf{x}_1)\chi_j(\mathbf{x}_1)r_{12}^{-1}\chi_k^*(\mathbf{x}_2)\chi_l(\mathbf{x}_2)d\mathbf{x}_1 d\mathbf{x}_2 \tag{2.26}$$

so that $\langle ij|kl\rangle = [ik|jl]$ and $[ij|kl] = \langle ik|jl\rangle$. The notation in (2.25) is often called physicists' notation and that in (2.26) is chemists' notation. Finally

$$\langle ij||kl\rangle \equiv \langle ij|kl\rangle - \langle ij|lk\rangle \tag{2.27}$$

Given $|\Psi_0\rangle = |\chi_1\chi_2\ldots\chi_N\rangle$, the energy E_0 is a functional of the spin orbitals $\{\chi_a\}$. We need to minimize $E_0[\{\chi_a\}]$ with respect to the spin orbitals subject to the spin orbitals being orthonormal. So the constraints are of the form

$$[a|b] - \delta_{ab} = 0$$

We will use the technique of Lagrange multipliers. Consider

$$\mathcal{L}[\{\chi_a\}] = E_0[\{\chi_a\}] - \sum_{a=1}^{N}\sum_{b=1}^{N}\varepsilon_{ba}([a|b] - \delta_{ab}) \tag{2.28}$$

where the ε_{ba} are the Lagrange multipliers. $\mathcal{L}$ is real and $[a|b] = [b|a]^*$ so that the ε_{ba} must be elements of a Hermitian matrix,

$$\varepsilon_{ba} = \varepsilon_{ab}^*$$

To get the minimum we need to minimize $\mathcal{L}$. We vary the spin orbitals by a small amount $\delta\chi_a$ and set the first variation in $\mathcal{L}$ equal to zero,

$$\delta\mathcal{L} = \delta E_0 - \sum_{a=1}^{N}\sum_{b=1}^{N}\varepsilon_{ba}\delta[a|b] = 0$$

It is easy to show that

$$\delta E_0 = \sum_{a=1}^{N}[\delta\chi_a|h|\chi_a] + \sum_{a=1}^{N}\sum_{b=1}^{N}[\delta\chi_a\chi_a|\chi_b\chi_b] - [\delta\chi_a\chi_b|\chi_b\chi_a] + \text{complex conjugate}$$

Now

$$\delta[a|b] = [\delta\chi_a|\chi_b] + [\chi_a|\delta\chi_b]$$

so that we can show that

$$\sum_{ab} \varepsilon_{ba}\delta[a|b] = \sum_{ab} \varepsilon_{ba}[\delta\chi_a|\chi_b] + \text{complex conjugate}$$

Therefore

$$\delta\mathcal{L} = \sum_{a=1}^{N}[\delta\chi_a|h|\chi_a] + \sum_{a=1}^{N}\sum_{b=1}^{N}[\delta\chi_a\chi_a|\chi_b\chi_b] - [\delta\chi_a\chi_b|\chi_b\chi_a]$$
$$-\sum_{a=1}^{N}\sum_{b=1}^{N}\varepsilon_{ba}[\delta\chi_a|\chi_b] + \text{complex conjugate } = 0 \tag{2.29}$$

2.2.2 The Coulomb and Exchange Operators

We now pause in our derivation to define and discuss some operators. The **coulomb operator** is

$$\mathcal{J}_b(1) \equiv \int |\chi_b(2)|^2 r_{12}^{-1} d\mathbf{x}_2 = \int \chi_b^*(2) r_{12}^{-1} \chi_b(2) d\mathbf{x}_2 \tag{2.30}$$

and it represents the average local potential at $\mathbf{x}_1$ arising from an electron in χ_b. We have electron-one in χ_a and electron-two in χ_b. Here we have used the shorthand notation $\chi_b(2) \equiv \chi_b(\mathbf{x}_2)$, and therefore the operator is defined in electron-one's coordinates. Clearly we do not lose any generality in doing this.

In exact theory the coulomb interaction is represented by the two-electron operator r_{ij}^{-1}. We replace this with a one-electron potential which we get by averaging the interaction r_{12}^{-1} of electron-one and electron-two over all space and spin coordinates $\mathbf{x}_2$ of electron-two weighted by the probability $|\chi_b(2)|^2 d\mathbf{x}_2$ that electron-two occupies $d\mathbf{x}_2$. If we sum over all $b \neq a$ we get the total average potential acting on the electron in χ_a.

Let $\mathcal{K}_b(1)$ be defined by

$$\mathcal{K}_b(1)\chi_a(1) = \left[\int \chi_b^*(2) r_{12}^{-1} \chi_a(2) d\mathbf{x}_2\right] \chi_b(1) \tag{2.31}$$

This is the **exchange operator**. By comparing this with the coulomb definition we see that if we operate with $\mathcal{K}_b(1)$ on $\chi_a(1)$ there is an *exchange* of electron-one and electron-two. This operator arises from the antisymmetric nature of the Slater determinant and does not have a simple interpretation. Unlike the coulomb operator the exchange operator is a nonlocal operator because there is no simple potential $\mathcal{K}_b(\mathbf{x}_1)$ defined at the point $\mathbf{x}_1$. When we operate with $\mathcal{K}_b(\mathbf{x}_1)$ on $\chi_a(\mathbf{x}_1)$ the value depends on χ_a over all space rather than just at $\mathbf{x}_1$.

If an electron is in χ_a the expectation values of $\mathcal{J}_b$ and $\mathcal{K}_b$ are the **coulomb** and **exchange integrals**, ie.

$$\langle\chi_a(1)|\mathcal{J}_b(1)|\chi_a(1)\rangle = \int \chi_a^*(1)\chi_a(1)r_{12}^{-1}\chi_b^*(2)\chi_b(2)d\mathbf{x}_1 d\mathbf{x}_2 = [aa|bb]$$

$$\langle\chi_a(1)|\mathcal{K}_b(1)|\chi_a(1)\rangle = \int \chi_a^*(1)\chi_b(1)r_{12}^{-1}\chi_b^*(2)\chi_a(2)d\mathbf{x}_1 d\mathbf{x}_2 = [ab|ba]$$

We now get back to our derivation.

2.2.3 A Transformation away from the Hartree-Fock Equations

We can use the definitions of the coulomb and exchange operators to write equation (2.29) as

$$\delta\mathcal{L} = \sum_{a=1}^{N}\int \delta\chi_a^*(1)\left[h(1)\chi_a(1) + \sum_{b=1}^{N}(\mathcal{J}_b(1) - \mathcal{K}_b(1))\chi_a(1) - \sum_{b=1}^{N}\varepsilon_{ba}\chi_b(1)\right]d\mathbf{x}_1 + \text{complex conjugate} = 0$$

Since $\delta\chi_a^*(1)$ is arbitrary $[\cdots] = 0$ for all a. Therefore

$$\left[h(1) + \sum_{b=1}^{N}\mathcal{J}_b(1) - \mathcal{K}_b(1)\right]\chi_a(1) = \sum_{b=1}^{N}\varepsilon_{ba}\,\chi_b(1) \qquad a = 1, 2, \ldots, N \tag{2.32}$$

The **(spin orbital) Fock operator** is defined by

$$f(1) \equiv h(1) + \sum_{b=1}^{N}\mathcal{J}_b(1) - \mathcal{K}_b(1) \tag{2.33}$$

When this acts on $\chi_a(1)$ the ath term in the sum vanishes

$$[\mathcal{J}_a(1) - \mathcal{K}_a(1)]\chi_a(1) = 0$$

This is obvious from the definitions given in equations (2.30) and (2.31). The Fock operator $f(1)$ is the sum of a core-Hamiltonian operator $h(1)$, and an effective one-electron potential operator called the **Hartree-Fock potential** $\nu^{\text{HF}}(1)$,

$$\nu^{\text{HF}}(1) \equiv \sum_{b=1}^{N}\mathcal{J}_b(1) - \mathcal{K}_b(1) \tag{2.34}$$

so that

$$f(1) = h(1) + \nu^{\text{HF}}(1)$$

Hence equation (2.32) is simply

$$f|\chi_a\rangle = \sum_{b=1}^{N}\varepsilon_{ba}|\chi_b\rangle \tag{2.35}$$

2.2.4 The Canonical Hartree-Fock Equations

We have not yet got it in the standard eigenvalue form of (2.22). The reason is that any single determinant $|\tilde{\Phi}\rangle$ formed from a set of spin orbitals $\{\chi_a\}$ keeps a degree of flexibility in the spin orbitals. For example we can cycle three columns of the determinant, or multiply two columns by -1. They can be mixed among themselves, by way of a unitary transformation, without changing $E_0 = \langle\tilde{\Phi}|\mathcal{H}_{\text{elec}}|\tilde{\Phi}\rangle$. This is shown below.

Now we need to consider unitary transformations of the spin orbitals among themselves. Let the new set $\{\chi'_a\}$ be obtained from $\{\chi_a\}$ by a unitary transformation

$$\chi'_a = \sum_b \chi_b U_{ba} \tag{2.36}$$

The transformation satisfies $U^H = U^{-1}$ and it preserves orthonormality. Define the matrix A by

$$A \equiv \begin{pmatrix} \chi_1(1) & \chi_2(1) & \cdots & \chi_N(1) \\ \chi_1(2) & \chi_2(2) & \cdots & \chi_N(2) \\ \vdots & \vdots & \ddots & \vdots \\ \chi_1(N) & \chi_2(N) & \cdots & \chi_N(N) \end{pmatrix} \tag{2.37}$$

so that the wave function $|\tilde{\Phi}\rangle$ is

$$|\tilde{\Phi}\rangle = (N!)^{-\frac{1}{2}} \det(A)$$

It is clear that

$$\begin{aligned} A' &= A\,U \\ &= \begin{pmatrix} \chi_1(1) & \chi_2(1) & \cdots & \chi_N(1) \\ \chi_1(2) & \chi_2(2) & \cdots & \chi_N(2) \\ \vdots & \vdots & & \vdots \\ \chi_1(N) & \chi_2(N) & \cdots & \chi_N(N) \end{pmatrix} \begin{pmatrix} U_{11} & U_{12} & \cdots & U_{1N} \\ U_{21} & U_{22} & \cdots & U_{2N} \\ \vdots & \vdots & & \vdots \\ U_{N1} & U_{N2} & \cdots & U_{NN} \end{pmatrix} \\ &= \begin{pmatrix} \chi'_1(1) & \chi'_2(1) & \cdots & \chi'_N(1) \\ \chi'_1(2) & \chi'_2(2) & \cdots & \chi'_N(2) \\ \vdots & \vdots & & \vdots \\ \chi'_1(N) & \chi'_2(N) & \cdots & \chi'_N(N) \end{pmatrix} \end{aligned}$$

Therefore $\det(A') = \det(U)\det(A)$ and

$$|\tilde{\Phi}'\rangle = \det(U)|\tilde{\Phi}\rangle$$

As U is unitary

$$\det(U) = \exp(i\phi)$$

and therefore $|\tilde{\Phi}'\rangle$ only differs from $|\tilde{\Phi}\rangle$ by a phase factor. If U is real then this is ± 1. Any observable property depends on $|\Psi|^2$ so the sign is unimportant. Thus we can consider the original wave function in terms of $\{\chi_a\}$ and the transformed wave function in terms of $\{\chi'_a\}$ to be identical. This means for a single determinant wave function any expectation value is invariant up to a unitary transformation of the spin orbitals. Therefore the spin orbitals that make the total energy stationary are not unique, and no particular physical significance can be given to a particular set of spin orbitals. Localized spin orbitals are no more "physical" than delocalized ones. By localized we mean spin orbitals that are like that shown in figure 3.3 on page 48.

We can use this to simplify (2.35) and put it in the standard eigenvalue form to get (2.22). To do this we need to consider the effect of unitary transformations on f and ε_{ab}. The only parts of the Fock operator that depend on the spin orbitals are the coulomb and exchange terms. It is straight forward to show that

$$\sum_a \mathcal{J}'_a(1) = \sum_b \mathcal{J}_b(1) \qquad \text{and} \qquad \sum_a \mathcal{K}'_a(1) = \sum_b \mathcal{K}_b(1)$$

so that the Fock operator is invariant under a unitary transformation of the spin orbitals,

$$f'(1) = f(1)$$

Multiplying (2.35) by $\langle\chi_c|$ shows that the Lagrange multipliers are matrix elements of the Fock operator

$$f_{ca} = \langle\chi_c|f|\chi_a\rangle = \sum_{b=1}^{N} \varepsilon_{ba}\langle\chi_c|\chi_b\rangle = \varepsilon_{ca}$$

By using this we can easily show that the new Lagrange multipliers are

$$\varepsilon' = U^H \varepsilon\, U$$

Since ε is Hermitian we can always choose U so that the above diagonalizes ε. We assume that the eigenvalues of ε are non-defective. Thus there exists a unique set $\{\chi'_a\}$ for which the matrix of Lagrange multipliers is diagonal, and

$$f|\chi'_a\rangle = \varepsilon'_a|\chi'_a\rangle$$

By dropping the primes we get the Hartree-Fock equation (2.22). The **canonical spin orbitals** are the solutions of this equation.

What we have done here is show that a unitary transformation of the spin orbitals does not change the energy of the Slater determinant. However a particular unitary transformation can be used to simplify the Hartree-Fock equations.

Note that the canonical spin orbitals will generally be delocalized. They will have certain symmetry properties characteristic of the symmetries of the molecule or of the Fock operator. Once we have got the canonical spin orbitals there are various ways of choosing a unitary transformation so that the transformed set of spin orbitals is in some sense localized. Having localized spin orbitals is in some sense more intuitive. If we have a molecule with two atoms say, then we would expect there to be an area between the two atoms where the probability or finding an electron is high. We would also expect this area to be associated with two electrons say, rather than all the electrons. This goes with the qualitative idea of a chemical bond.

2.2.5 Koopmans' Theorem

This theorem gives an interpretation of the eigenvalues of the Hartree-Fock equation. It was due to Koopmans in 1933 [18]. We will state the theorem and then explain it. This theorem will not be used for anything later on but it is important to see what the eigenvalues mean. Also by explaining the theorem some notation that we will be using later on can be defined.

When we have solved the Hartree-Fock equation for the N occupied spin orbitals $\{\chi_a\}$ the Fock operator, which depends on these orbitals, becomes a well defined Hermitian operator. It will have an infinite number of eigenfunctions

$$f|\chi_j\rangle = \varepsilon_j|\chi_j\rangle \qquad\qquad j = 1, 2, \ldots$$

The N spin orbitals with the lowest energies are the ones that are **occupied** in $|\Psi_0\rangle$ and we label these with indices $a, b, \ldots$. The other spin orbitals are **virtual** or **unoccupied**. There are an infinite number of these and they are labelled with $r, s, \ldots$. The occupied and virtual orbitals are shown in figure 2.2 with the ground state occupancy.

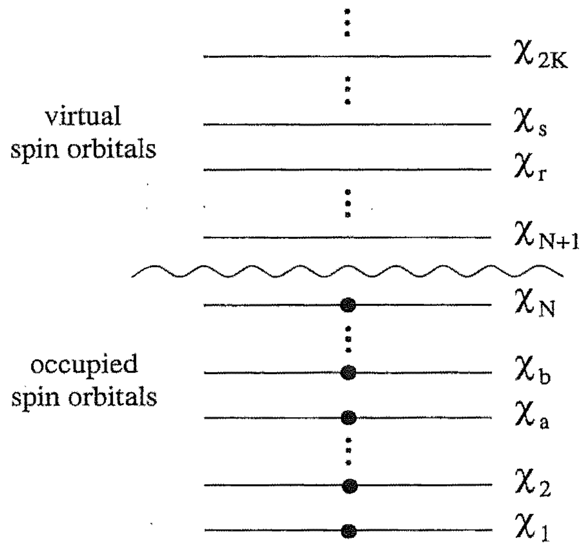


Figure 2.2: The Hartree-Fock ground state $|^{N}\Psi_0\rangle$.

Let $|^{N}\Psi\rangle$ be an N–electron Slater determinant.

Koopmans' Theorem: *Let $|^{N}\Psi_0\rangle$ have occupied and virtual spin orbital energies ε_a and ε_r respectively. The ionization potential to produce $|^{N\text{-}1}\Psi_a\rangle$ with identical spin orbitals, obtained by removing an electron from spin orbital χ_a, is just $-\varepsilon_a$. The electron affinity to produce $|^{N+1}\Psi^r\rangle$ with identical spin orbitals, obtained by adding an electron to spin orbital χ_r, is $-\varepsilon_r$.*

It is easy to show that

$$\varepsilon_a = \langle a|h|a\rangle + \sum_{b\neq a}\langle ab|ab\rangle - \langle ab|ba\rangle$$

and

$$\varepsilon_r = \langle r|h|r\rangle + \sum_{b=1}^{N}\langle rb|rb\rangle - \langle rb|br\rangle$$

So ε_a is the sum of kinetic energy and attraction to the nuclei $\langle a|h|a\rangle$, and coulomb $\langle ab|ab\rangle$ and exchange $-\langle ab|ba\rangle$ interactions with the other $N-1$ electrons in the spin orbitals $|\chi_b\rangle$, $b \neq a$. Note that $\langle ab|ba\rangle \neq 0$ only if the spins of the electrons in $|\chi_a\rangle$ and $|\chi_b\rangle$ are parallel. The value of ε_r has the same interpretation, except it includes interactions with all N electrons of $|\Psi_0\rangle$. An electron has been added to $|\Psi_0\rangle$ to give an $(N+1)$–electron state and ε_r represents the electron's energy.

Now the total energy of the N–electron system is

$$E_0 = \sum_{a=1}^{N} \langle a|h|a\rangle + \tfrac{1}{2}\sum_{a=1}^{N}\sum_{b=1}^{N} \langle ab||ab\rangle \neq \sum_{a=1}^{N} \varepsilon_a$$

The half is there so that the interactions are not calculated twice. The equation says that the orbital energies do not sum to give the total energy.

Now suppose we remove an electron from χ_c. Put

$$|{}^{\mathrm{N}}\Psi_0\rangle \equiv |\chi_1\chi_2\cdots\chi_c\cdots\chi_N\rangle \tag{2.38}$$

and

$$|{}^{\mathrm{N\text{-}1}}\Psi_c\rangle \equiv |\chi_1\chi_2\cdots\chi_{c-1}\chi_{c+1}\cdots\chi_N\rangle \tag{2.39}$$

We can show that the **ionization potential** for this process is

$$\mathrm{IP} \equiv {}^{N-1}E_c - {}^{N}E_0 = -\varepsilon_c \tag{2.40}$$

Depending on which orbital we remove an electron from the state $|{}^{\mathrm{N\text{-}1}}\Psi_c\rangle$ may or may not represent the ground state of the ionized species. We cannot expect the optimum spin orbitals of $|{}^{\mathrm{N\text{-}1}}\Psi_c\rangle$ to be the same as those of $|{}^{\mathrm{N}}\Psi_0\rangle$. When we move an electron the ones that are left will "rearrange" themselves so that their energy is lower. This is because of the entropy principle. Here we have assumed that the optimum orbitals are the same. The orbital energies represent the energy needed to remove an electron from the spin orbital. Orbital energies are usually negative and ionization potentials are positive.

Now if we add an electron to one of the unoccupied spin orbitals χ_r to produce the $(N+1)$–electron single determinant

$$|{}^{\mathrm{N+1}}\Psi^r\rangle \equiv |\chi_r\chi_1\chi_2\cdots\chi_N\rangle \tag{2.41}$$

we can show that the **electron affinity** for this process is

$$\mathrm{EA} \equiv {}^{N}E_0 - {}^{N+1}E^r = -\varepsilon_r \tag{2.42}$$

Again we have ignored the fact that the optimum spin orbitals of the new species will probably be different. If ε_r is negative, which corresponds to $|{}^{\mathrm{N+1}}\Psi^r\rangle$ being more stable than $|{}^{\mathrm{N}}\Psi_0\rangle$ the EA is positive.

So basically Koopmans' theorem gives a way of calculating approximate IP's and EA's. It is a *frozen orbital* approximation. ${}^{N-1}E_a$ and ${}^{N+1}E^r$ will really be lower than the values above so that we get IP's that are too positive and EA's that are too negative. Generally Koopmans IP's are good first approximations to experimental IP's but the EA's are often bad. This is because the correlation effects the the Hartree-Fock approximation ignores give energies that cancel the error in IP's but add to it in EA's.

Now we mention some other notation. If an electron is excited from χ_a to χ_r we use the notation

$$|^{N}\Psi_a^r\rangle = |\Psi_a^r\rangle \equiv |\chi_1\chi_2 \cdots \chi_r\chi_b \cdots \chi_N\rangle \tag{2.43}$$

This situation is shown in figure 2.3.

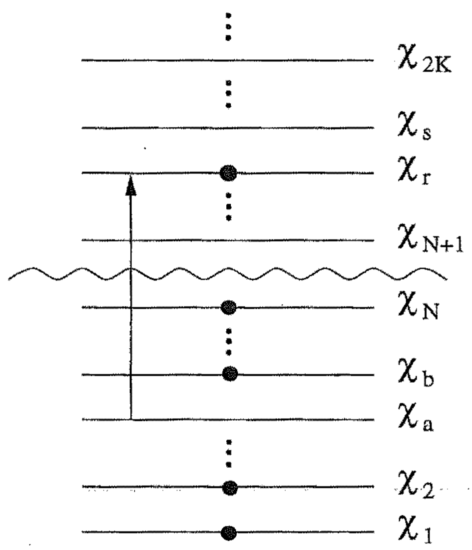


Figure 2.3: A singly excited determinant $|^{N}\Psi_a^r\rangle$.

2.2.6 Brillouin's Theorem

We now state another theorem which is important when looking at how different configurations interact. This theorem will not be proved. The result will be needed later on.

Brillouin's Theorem: *Singly excited determinants $|\Psi_a^r\rangle$ will not interact directly with a reference Hartree-Fock determinant $|\Psi_0\rangle$, that is*

$$\langle\Psi_0|\mathcal{H}_{elec}|\Psi_a^r\rangle = 0$$

The exact wave function for any state of the system can be expressed as

$$|\Phi\rangle = c_0|\Psi_0\rangle + \sum_{ra} c_a^r|\Psi_a^r\rangle + \sum_{\substack{a<b \\ r<s}} c_{ab}^{rs}|\Psi_{ab}^{rs}\rangle + \cdots$$

So we might expect the singly excited determinants to give the leading correction to $|\Psi_0\rangle$. However Brillouin's theorem shows that the Hartree-Fock ground state cannot be improved by mixing it with singly excited determinants. This kind of idea will be discussed more in chapter 4 where we look at going beyond the Hartree-Fock approximation.

Chapter 3

The Self-Consistent Field Procedure

We now look at how the Hartree-Fock equations can be converted into numerical equations that we can solve. These equations apply in the restricted closed-shell case and are called the Roothaan equations. To get them it is necessary to introduce a set of known basis functions.

In the section 3.2 the self-consistent field procedure is looked at. It is the most simple iterative method that gets used to solve the nonlinear Roothaan equations. Although the method is rather dated, it brings up a lot of ideas that will be important in later chapters.

In the last section of this chapter we look at the types of basis functions that get used. An appropriate choice of basis functions is important because the results depend greatly on the basis set.

3.1 The Roothaan Equations

Before we can go any further in solving the Hartree-Fock equations we need to be more specific about the form of the spin orbitals. Before doing this we need to define a few more terms. After that spin is eliminated from the equations, and a basis set is introduced. The basis set is used to convert the closed shell Hartree-Fock equations into a set of algebraic equations, the Roothaan equations.

3.1.1 Restricted Closed-Shell Wave Functions

Given a set of N orthonormal spatial orbitals we can form a set of $2N$ spin orbitals by multiplying each spatial orbital by either α or β spin function. Such spin orbitals are **restricted spin orbitals**, and determinants formed from them are called **restricted determinants**. In a determinant each ψ_i can be occupied by one or two electrons. If each spatial orbital is doubly occupied then we have a **closed-shell determinant**. An **open-shell** is a spatial orbital that contains a single electron. An **open-shell determinant** involves an open-shell spatial orbital, and they get referred to by the number of open-shells they involve. When atoms or molecules have partially filled sub-shells we find we can get a slightly lower variational energy if paired electrons are allowed to have different spatial orbitals. This gives an **unrestricted** Hartree-Fock wave function. In the closed-shell case the SCF wave function can be written as a single Slater determinant, but in the open-shell case it is necessary to write the wave function as a linear combination of Slater determinants.

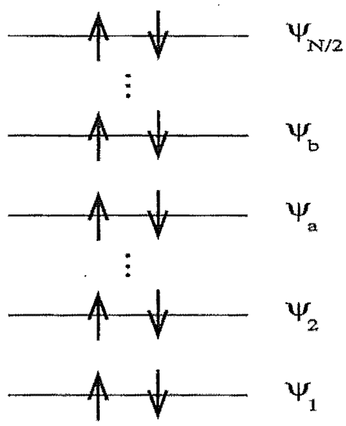


Figure 3.1: A closed-shell restricted Hartree-Fock ground state determinant $|\Psi_0\rangle$ with N electrons.

Here we are only concerned with restricted Hartree-Fock wave functions and specifically closed-shell calculations. This means the molecular states are only allowed to have an even number of electrons, with all electrons paired so that $n = N/2$ spatial orbitals are doubly occupied. This is shown in figure 3.1. We are looking at closed-shell *ground* states. In principle what we are talking about could be used for calculating excited states. However if we were trying to calculate an excited state we might not have a way of stopping it from converging to the ground state, this is the case when the excited and ground states have the same symmetry.

A restricted set of spin orbitals has the form

$$\chi_{2j-1}(\mathbf{x}) = \psi_j(\mathbf{r})\alpha(\omega) \equiv \psi_j$$
$$\chi_{2j}(\mathbf{x}) = \psi_j(\mathbf{r})\beta(\omega) \equiv \overline{\psi}_j$$

where $j = 1, 2, \ldots, \frac{N}{2}$, and the closed-shell restricted ground state is

$$|\Psi_0\rangle = |\chi_1\chi_2\cdots\chi_{N-1}\chi_N\rangle \equiv |\psi_1\overline{\psi}_1\cdots\psi_a\overline{\psi}_a\cdots\psi_{\frac{N}{2}}\overline{\psi}_{\frac{N}{2}}\rangle \tag{3.1}$$

The **restricted Hartree-Fock** (RHF) method is used for the calculation of open-shell SCF wave functions. A method for this was devised by Roothaan in 1960 [30].

In later chapters a lot of the stuff we talk about will apply to unrestricted and open-shell wave functions too. However the actual equations can get more complicated. Unrestricted open-shell Hartree-Fock wave functions and the resulting Pople-Nesbet equations are looked at in [42, pages 205–229]. The Hartree-Fock equations as discussed in the previous chapter apply in these more general situations.

3.1.2 Elimination of Spin

We want to convert the general spin orbital Hartree-Fock equations into a spatial eigenvalue equation with each of the spatial molecular orbitals occupied twice. We get this by integrating over the spin functions.

The spin orbital $\chi_i(\mathbf{x}_1)$ will have either α or β spin function. Without loss of generality assume it has spin function α, so that

$$f(\mathbf{x}_1)\psi_j(\mathbf{r}_1)\alpha(\omega_1) = \varepsilon_j\psi_j(\mathbf{r}_1)\alpha(\omega_1)$$

where

$$\varepsilon_j \;\equiv\; \text{energy of spatial orbital } \psi_j \quad = \quad \varepsilon_i \;\equiv\; \text{energy of spin orbital } \chi_i$$

Multiplying on the left by $\alpha^*(\omega_1)$ and integrating over the spin gives

$$\left[\int \alpha^*(\omega_1)f(\mathbf{x}_1)\alpha(\omega_1)d\omega_1\right]\psi_j(\mathbf{r}_1) = \varepsilon_j\psi_j(\mathbf{r}_1) \tag{3.2}$$

Let $f(\mathbf{r}_1)$ be the **closed-shell Fock operator**

$$f(\mathbf{r}_1) \equiv \int \alpha^*(\omega_1) f(\mathbf{x}_1) \alpha(\omega_1) d\omega_1 \tag{3.3}$$

We can write the spin orbital Fock operator as

$$f(\mathbf{x}_1) = h(\mathbf{r}_1) + \sum_{c=1}^{N} \int \chi_c^*(\mathbf{x}_2) r_{12}^{-1} (1 - \mathcal{P}_{12}) \chi_c(\mathbf{x}_2) d\mathbf{x}_2$$

where $\mathcal{P}_{12}$ is an operator that interchanges the coordinates of electron-one and electron-two. Using the two equations above (3.2) becomes

$$\begin{aligned} f(\mathbf{r}_1)\psi_j(\mathbf{r}_1) &= h(\mathbf{r}_1)\psi_j(\mathbf{r}_1) + \sum_c \int \alpha^*(\omega_1)\chi_c^*(\mathbf{x}_2) r_{12}^{-1} \chi_c(\mathbf{x}_2)\alpha(\omega_1) d\omega_1 d\mathbf{x}_2 \psi_j(\mathbf{r}_1) \\ &\quad - \sum_c \int \alpha^*(\omega_1)\chi_c^*(\mathbf{x}_2) r_{12}^{-1} \chi_c(\mathbf{x}_1)\alpha(\omega_2) d\omega_1 d\mathbf{x}_2 \psi_j(\mathbf{r}_2) \\ &= \varepsilon_j \psi_j(\mathbf{r}_1) \end{aligned}$$

Since we have a closed shell

$$\sum_{c=1}^{N} \rightarrow \sum_{c=1}^{N/2} + \sum_{c=1}^{N/2}$$

and after integrating over ω_1 and ω_2 and simplifying we get

$$\begin{aligned} f(\mathbf{r}_1)\psi_j(\mathbf{r}_1) &= h(\mathbf{r}_1)\psi_j(\mathbf{r}_1) + \left[2\sum_{c=1}^{N/2} \int \psi_c^*(\mathbf{r}_2) r_{12}^{-1} \psi_c(\mathbf{r}_2) d\mathbf{r}_2 \right] \psi_j(\mathbf{r}_1) \\ &\quad - \sum_{c=1}^{N/2} \left[\int \psi_c^*(\mathbf{r}_2) r_{12}^{-1} \psi_j(\mathbf{r}_2) d\mathbf{r}_2 \right] \psi_c(\mathbf{r}_1) \\ &= \varepsilon_j \psi_j(\mathbf{r}_1) \end{aligned}$$

Hence the closed-shell Fock operator has the form

$$f(\mathbf{r}_1) = h(\mathbf{r}_1) + \sum_{a=1}^{N/2} \int \psi_a^*(\mathbf{r}_2) r_{12}^{-1} (2 - \mathcal{P}_{12}) \psi_a(\mathbf{r}_2)$$

or equivalently

$$f(1) = h(1) + \sum_{a=1}^{N/2} 2J_a(1) - K_a(1) \tag{3.4}$$

where we have defined the **closed-shell coulomb** and **exchange operators** by

$$J_a(1) \equiv \int \psi_a^*(2) r_{12}^{-1} \psi_a(2) d\mathbf{r}_2 \tag{3.5}$$

$$K_a(1)\psi_i(1) \equiv \left[\int \psi_a^*(2) r_{12}^{-1} \psi_i(2) d\mathbf{r}_2 \right] \psi_a(1) \tag{3.6}$$

These are analogous to those for spin orbitals except for the factor of 2. There is no factor of 2 on the other term in the sum because there is only an exchange interaction between electrons of parallel spins. The **closed-shell spatial Hartree-Fock equation** is just

$$f(1)\psi_j(1) = \varepsilon_j \psi_j(1)$$

For the closed-shell determinant $|\Psi_0\rangle = |\psi_1\overline{\psi}_1 \cdots \psi_{\frac{N}{2}}\overline{\psi}_{\frac{N}{2}}\rangle$ the Hartree-Fock energy is

$$\begin{aligned} E_0 &= \langle\Psi_0|\mathcal{H}_{\text{elec}}|\Psi_0\rangle = 2\sum_a (a|h|a) + \sum_a\sum_b 2(aa|bb) - (ab|ba) \\ &= 2\sum_a h_{aa} + \sum_a\sum_b 2J_{ab} - K_{ab} \end{aligned} \tag{3.7}$$

where we have used the *spatial orbital* notation

$$(i|h|j) = h_{ij} = (\psi_i|h|\psi_j) \equiv \int \psi_i^*(\mathbf{r}_1)h(\mathbf{r}_1)\psi_j(\mathbf{r}_1)d\mathbf{r}_1 \tag{3.8}$$

$$(ij|kl) = (\psi_i\psi_j|\psi_k\psi_l) \equiv \int \psi_i^*(\mathbf{r}_1)\psi_j(\mathbf{r}_1)r_{12}^{-1}\psi_k^*(\mathbf{r}_2)\psi_l(\mathbf{r}_2)d\mathbf{r}_1 d\mathbf{r}_2 \tag{3.9}$$

$$J_{ij} \equiv (ii|jj) \qquad \text{(coulomb integral)} \tag{3.10}$$

$$K_{ij} \equiv (ij|ji) \qquad \text{(exchange integral)} \tag{3.11}$$

In the real case the integrals $(ij|kl)$ are invariant with respect to the interchange of the labels $i \leftrightarrow j$, $k \leftrightarrow l$ and $ij \leftrightarrow kl$.

The closed-shell spatial orbital form of the orbital energies is

$$\varepsilon_i = (\psi_i|h|\psi_i) + \sum_{b=1}^{N/2} 2(ii|bb) - (ib|bi) = h_{ii} + \sum_{b=1}^{N/2} 2J_{ib} - K_{ib} \tag{3.12}$$

3.1.3 Getting the Roothaan Equations

Now that we no longer need to worry about spin, calculating the molecular orbitals is equivalent to solving

$$f(\mathbf{r}_1)\psi_i(\mathbf{r}_1) = \varepsilon_i\psi_i(\mathbf{r}_1) \tag{3.13}$$

We could solve this numerically for atoms, but we have not got any practical procedure for getting numerical solutions for molecules. Roothaan [31] and Hall [12] more or less simultaneously showed that by introducing a set of *known* spatial basis functions the differential equation could be converted to a set of algebraic equations. The resulting equations can be solved by standard matrix techniques. These basis functions will be called atomic orbitals in later chapters. This is because they are usually the sort of functions that describe the orbitals in an atom.

What we do is introduce a set of N_{BF} known basis functions $\{\phi_\mu(\mathbf{r}) : \mu = 1, 2, \ldots, N_{\mathrm{BF}}\}$ and expand the unknown molecular orbitals in these basis functions,

$$\psi_i = \sum_{\mu=1}^{N_{\mathrm{BF}}} C_{\mu i}\phi_\mu \qquad\qquad i = 1, 2, \ldots, N_{\mathrm{BF}} \tag{3.14}$$

Note that if we put the functions into vectors they will be row vectors, and we have the relationship

$$\boldsymbol{\psi} = \boldsymbol{\phi} C$$

The vectors $\boldsymbol{\psi}$ and $\boldsymbol{\phi}$ contain the molecular orbitals and basis functions respectively, and C is an $N_{\mathrm{BF}} \times N_{\mathrm{BF}}$ matrix consisting of the expansion coefficients. If $\{\phi_\mu\}$ was complete this would be exact, but in order for it to be complete it needs to be infinite and we obviously always need to have a finite set of N_{BF} basis functions for practical reasons. We need to choose a basis that will provide a reasonably accurate expansion for the exact molecular orbitals. In particular we would like it to be accurate for the molecular orbitals that are occupied in $|\Psi_0\rangle$ and determine the ground state energy E_0.

Using a basis set of N_{BF} functions we get a set of $2N_{\mathrm{BF}}$ spin orbitals, N_{BF} with α spin and N_{BF} with β spin. N of these are occupied so that $2N_{\mathrm{BF}} - N$ are virtual spin orbitals.

Later we will look at how we go about choosing the basis and what sort of functions get used as basis functions. For the moment we will just assume it is known.

As the basis set becomes more complete equation (3.14) gives more accurate representations of the exact molecular orbitals, and these molecular orbitals converge to those of equation (3.13), which are the eigenfunctions of the Fock operator.

From (3.14) we see that the problem of calculating the Hartree-Fock molecular orbitals reduces to the problem of calculating the set of expansion coefficients $C_{\mu i}$. Now we will look at how to put these into a matrix equation. By substituting equation (3.14) into equation (3.13), multiplying by $\phi_\mu^*(1)$ on the left, and integrating we get

$$\sum_\nu C_{\nu i} \int \phi_\mu^*(1) f(1) \phi_\nu(1) d\mathbf{r}_1 = \varepsilon_i \sum_\nu C_{\nu i} \int \phi_\mu^*(1)\phi_\nu(1) d\mathbf{r}_1 \tag{3.15}$$

We now need to define a couple of matrices. The **overlap matrix** S is given by

$$S_{\mu\nu} \equiv \int \phi_\mu^*(1)\phi_\nu(1)d\mathbf{r}_1 \tag{3.16}$$

and it is $N_{\mathrm{BF}} \times N_{\mathrm{BF}}$ and hermitian. However it is usually real and symmetric. We assume that the basis functions are normalized and linearly independent, but not necessarily orthogonal. Therefore the overlap has magnitude $0 \leq |S_{\mu\nu}| \leq 1$ by the Cauchy-Schwarz inequality. The diagonal elements are unity and the off-diagonal elements are numbers less than one in magnitude. The sign of the off-diagonal elements depends on the relative sign of the two basis functions and their relative orientation and separation in space. If two off-diagonal elements approach unity in magnitude, which is complete overlap, then the two basis functions approach linear dependence. S can be diagonalized by a unitary matrix. The eigenvalues of S can be shown to be positive so that S is positive definite. Linear dependence in the basis set occurs when $\det S = 0$. Sometimes S is called the metric matrix.

The **Fock matrix** F has elements

$$F_{\mu\nu} \equiv \int \phi_\mu^*(1)f(1)\phi_\nu(1)d\mathbf{r}_1 \tag{3.17}$$

and it is also $N_{\mathrm{BF}} \times N_{\mathrm{BF}}$ and hermitian, but usually real and symmetric. The Fock operator $f(1)$ is a one-electron operator and any set of one-electron functions will define a matrix representation of this operator. So F is the matrix representation of the Fock operator with the set of basis functions $\{\phi_\mu\}$.

Now in terms of the Fock matrix and overlap matrix equation (3.15) is

$$\sum_\nu F_{\mu\nu}C_{\nu i} = \varepsilon_i \sum_\nu S_{\mu\nu}C_{\nu i} \qquad i = 1, \ldots, N_{\mathrm{BF}}. \tag{3.18}$$

These are the **Roothaan equations** and we can write them in matrix form as

$$F\,C = S\,C\,\varepsilon \tag{3.19}$$

where C consists of the expansion coefficients $C_{\mu i}$. Note that the coefficients describing ψ_i are in the ith column of C. The matrix ε is diagonal and consists of the orbital energies ε_i,

$$\varepsilon = \begin{pmatrix} \varepsilon_1 & & 0 \\ & \ddots & \\ 0 & & \varepsilon_{N_{\mathrm{BF}}} \end{pmatrix} \tag{3.20}$$

Using the fact that the molecular orbitals $\{\psi_i\}$ are orthonormal it is easy to show that

$$C^H S\, C = I$$

3.2 The Self-Consistent Field Procedure

The **self-consistent field** (SCF) procedure is the basic method for solving the Roothaan equations. In this section it is described in its most naive form. Later in chapter 5 we will look at how this method can be improved. However in order to see how the method works it is helpful to look at the basic idea.

3.2.1 The Density Matrix

At this point the problem of determining the Hartree-Fock molecular orbitals $\{\psi_i\}$ and orbital energies ε_i involves solving the equation $FC = SC\varepsilon$. To continue, we need an explicit expression for the Fock matrix. But before we can do this we need to introduce the concept of a density matrix.

If an electron is described by $\psi_a(\mathbf{r})$ then the probability of finding the electron in $d\mathbf{r}$ at $\mathbf{r}$ is $|\psi_a(\mathbf{r})|^2 d\mathbf{r}$. So the probability distribution, or charge density, is just $|\psi_a(\mathbf{r})|^2$. For a closed-shell molecule described by a single determinant wave function the **total charge density** is

$$\rho(\mathbf{r}) = 2\sum_{a=1}^{N/2} |\psi_a(\mathbf{r})|^2 \tag{3.21}$$

The total number of electrons is

$$\int \rho(\mathbf{r})d\mathbf{r} \;=\; 2\sum_{a=1}^{N/2} \int |\psi_a(\mathbf{r})|^2 d\mathbf{r} \;=\; 2\sum_{a=1}^{N/2} 1 \;=\; N$$

So the probability of finding any electron in $d\mathbf{r}$ at $\mathbf{r}$ is given by $N^{-1}\rho(\mathbf{r})d\mathbf{r}$. Substituting (3.14) into (3.21) gives

$$\begin{aligned}
\rho(\mathbf{r}) &= 2\sum_{a=1}^{N/2} \left(\sum_{\nu} C^*_{\nu a}\phi^*_\nu(\mathbf{r})\right)\left(\sum_{\mu} C_{\mu a}\phi_\mu(\mathbf{r})\right) \\
&= \sum_{\mu\nu} \left[2\sum_{a=1}^{N/2} C_{\mu a}C^*_{\nu a}\right] \phi_\mu(\mathbf{r})\phi^*_\nu(\mathbf{r}) \\
&= \sum_{\mu\nu} D_{\mu\nu}\phi_\mu(\mathbf{r})\phi^*_\nu(\mathbf{r}) \qquad (3.22)
\end{aligned}$$

where we have defined a **density matrix** D by

$$D_{\mu\nu} \equiv 2 \sum_{a=1}^{N/2} C_{\mu a} C_{\nu a}^{*} \tag{3.23}$$

Therefore $D = 2CC^H$ and D specifies $\rho(\mathbf{r})$ completely when we have the set of known basis functions. The notation P often gets used instead of D. This is because P is considered to be a capital ρ and the matrix can be thought of as a generalization of the total charge density ρ.

It is easy to show that the Fock operator can be expressed as

$$f(\mathbf{r}_1) = h(\mathbf{r}_1) + \tfrac{1}{2} \sum_{\lambda\sigma} D_{\lambda\sigma} \left[\int \phi_{\sigma}^{*}(\mathbf{r}_2) \left(2 - \mathcal{P}_{12}\right) r_{12}^{-1} \phi_{\lambda}(\mathbf{r}_2) d\mathbf{r}_2 \right] \tag{3.24}$$

3.2.2 Description of the SCF Procedure

The self-consistent field procedure is the computational method used to get restricted closed-shell Hartree-Fock wave functions for molecules. That is, it is used to obtain $|\Psi_0\rangle$. It will be described fully in subsection 3.2.5. We now outline it.

The basic idea behind the SCF method is to make an initial guess at the spatial orbitals, which we do by guessing at the density matrix, and then solve $f(i)\psi_i(1) = \varepsilon_i \psi_i(1)$ for a new set of spatial orbitals. From the new states we get the new density matrix, then we get a new field f, and we repeat the procedure. When self-consistency is reached, which is when the fields no longer change and the spin orbitals used to calculate the Fock operator are the same as its eigenfunctions, we stop. This is why the Hartree-Fock equations are often called the self-consistent field equations.

The actual iterative scheme that is used can be expressed as

$$F[C^{(i)}]\, C^{(i+1)} = S\, C^{(i+1)}\, \varepsilon^{(i+1)}$$

We are using a fixed point iterative scheme and solving a generalized eigenvalue equation at each iteration. Often the Fock matrices do not change much between iterations and the convergence can be slow. We will look at improving convergence later.

3.2.3 The Fock Matrix

As we have already said an explicit expression for the Fock matrix F is needed. By using expression (3.24) we get

$$\begin{aligned} F_{\mu\nu} &= \int \phi_\mu^*(1) f(1) \phi_\nu(1) d\mathbf{r}_1 \\ &= \int \phi_\mu^*(1) h(1) \phi_\nu(1) d\mathbf{r}_1 + \sum_{\lambda\sigma} D_{\lambda\sigma} \left[(\mu\nu|\sigma\lambda) - \tfrac{1}{2}(\mu\lambda|\sigma\nu) \right] \\ &= H_{\mu\nu}^{\text{core}} + G_{\mu\nu} \end{aligned} \tag{3.25}$$

Here we have defined a **core-Hamiltonian matrix,**

$$H_{\mu\nu}^{\text{core}} \equiv \int \phi_\mu^*(1) h(1) \phi_\nu(1) d\mathbf{r}_1 \tag{3.26}$$

Recall that the one-electron operator $h(1)$ describes the kinetic energy and nuclear attraction of an electron

$$h(1) = -\tfrac{1}{2}\nabla_1^2 - \sum_A \frac{Z_A}{|\mathbf{r}_1 - \mathbf{R}_A|}$$

To calculate the core-Hamiltonian matrix elements we have to work out the **kinetic energy integrals**

$$T_{\mu\nu} \equiv \int \phi_\mu^*(1) \left[-\tfrac{1}{2}\nabla_1^2 \right] \phi_\nu(1) d\mathbf{r}_1 \tag{3.27}$$

and the **nuclear attraction integrals**

$$V_{\mu\nu}^{\text{nucl}} \equiv \int \phi_\mu^*(1) \left[-\sum_A \frac{Z_A}{|\mathbf{r}_1 - \mathbf{R}_A|} \right] \phi_\nu(1) d\mathbf{r}_1 \tag{3.28}$$

with

$$H_{\mu\nu}^{\text{core}} = T_{\mu\nu} + V_{\mu\nu}^{\text{nucl}}$$

Note that the core-Hamiltonian matrix only has to be calculated once unlike the rest of the Fock matrix.

We have also defined the matrix G and it is the two-electron part of the Fock matrix which has to be recalculated at each iteration. It depends on D and the set of two-electron basis function integrals $(\mu\nu|\lambda\sigma)$. There are N_{BF}^4 of these integrals and in the real case $\sim \frac{1}{8} N_{\text{BF}}^4$ of them are unique [16, page 22]. Recall from the introduction that N_{BF} tends to range from $O(10^2)$ to $O(10^3)$. This means that if, in the real case, we had 5000 basis functions we would have $O(10^{14})$ unique two-electron integrals. So calculating these integrals can be a *very* big task.

Later on we will need to express the Fock matrix in terms of the $\{\psi_i\}$ as well.

3.2.4 Orthogonalizing the Basis

The Roothaan equations are nonlinear and can be expressed as $F(C)C = SC\varepsilon$. If we had an orthonormal basis set this would just be $FC = C\varepsilon$ and we could find the eigenvectors in C and the eigenvalues in ε by diagonalizing F. So in order to put Roothaan's equations into the usual eigenvalue form we need to consider orthogonalizing the basis functions.

We can always find a transformation matrix X such that the transformed set $\{\phi'_\mu\}$ given by

$$\phi'_\mu = \sum_\nu X_{\nu\mu}\phi_\nu \qquad \mu = 1, 2, \ldots, N_{\text{BF}} \tag{3.29}$$

satisfy

$$\int \phi'^*_\mu(\mathbf{r})\phi'_\nu(\mathbf{r})d\mathbf{r} = \delta_{\mu\nu}$$

It is easy to show that this means we need to have

$$X^H S X = I$$

Since S is Hermitian there exists a diagonal s and a unitary U such that

$$U^H S U = s$$

There are two common ways of orthogonalizing the basis $\{\phi_\mu\}$. The first is **symmetric orthogonalization**. Let

$$X \equiv S^{-\frac{1}{2}} = U s^{-\frac{1}{2}} U^H \tag{3.30}$$

By substituting this into $X^H S X = I$ we can see that this choice works. The eigenvalues of S are all positive so there is no problem in taking the square root. If the basis is nearly linearly dependent then some of the eigenvalues will approach zero and the above will involve dividing by quantities that are nearly zero. So this way of orthogonalizing will lead to problems in numerical stability if the basis set is nearly linear dependent.

Now the second way is **canonical orthogonalization** which uses

$$X \equiv U s^{-\frac{1}{2}} \tag{3.31}$$

The columns of U are divided by the square root of the corresponding eigenvalue, $X_{ij} = U_{ij}/s_j^{\frac{1}{2}}$. By substituting this into $X^H S X = I$ we can see that it works. It seems that there will again be problems if there is linear dependence in the basis. However we can order the eigenvalues in s any way we like as long as U is reordered too. Suppose $s_1 > s_2 > s_3 > \ldots$ and that the last m of these are small enough to give numerical problems. We can truncate the transformation matrix by chopping off its last m columns to give $\widetilde{X}$ with $\widetilde{X}_{ij} = U_{ij}/s_j^{\frac{1}{2}}$. So we only get $N_{\text{BF}} - m$ transformed orthonormal basis functions

$$\phi'_\mu = \sum_{\nu=1}^{N} \widetilde{X}_{\nu\mu}\phi_\nu \qquad\qquad \mu = 1, 2, \ldots, N_{\text{BF}} - m \tag{3.32}$$

We have thrown away part of the basis set.

One way of dealing with the problem of a nonorthogonal basis is to get $\{\phi'_\mu\}$ and work with it through out the calculations. This would eliminate S from Roothaan's equation which could then be solved by diagonalization. We would have to calculate all the two-integrals using the new orbitals, or transform the $(\mu\nu|\lambda\sigma)$ into $(\mu'\nu'|\lambda'\sigma')$. This is time consuming so we do it a more efficient way.

Consider the new coefficient matrix C' given by

$$C' = X^{-1}C$$

We have assumed that X^{-1} exists and this will be true if we have linear independence. Putting $C = XC'$ into the Roothaan equations and multiplying by X^H gives

$$(X^H F X)C' = (X^H S X)C'\varepsilon$$

and so

$$F'C' = C'\varepsilon \tag{3.33}$$

where F' is defined by

$$F' = X^H F X$$

Equation (3.33) is the **transformed Roothaan equation** and we can solve it for C' by diagonalizing F'. Note that the equation is still nonlinear.

In the next subsection we summarise the method above.

3.2.5 The SCF Method

Before we can apply the SCF method to a molecule we obviously need to have a basis set. This will be looked at in the next section. We will also assume that we have a way of getting an initial guess at the density matrix.

The **self-consistent field procedure** is:

(1) Choose a basis set $\{\phi_\mu\}$, and a molecule, by stating a set of nuclear coordinates $\{\mathbf{R}_A\}$, a set of atomic numbers $\{Z_A\}$ and the number of electrons N.

(2) Calculate all the molecular integrals that are needed, $S_{\mu\nu}$, $H_{\mu\nu}^{\text{core}}$, and $(\mu\nu|\lambda\sigma)$.

(3) Diagonalize the overlap matrix S and get a transformation matrix X from either (3.30) or (3.31).

(4) Work out a guess at the density matrix D of equation (3.23).

(5) Using the density matrix D and the two-electron integrals $(\mu\nu|\lambda\sigma)$ calculate the matrix G of equation (3.25).

(6) Add the matrix G to the core-Hamiltonian matrix to obtain the Fock matrix $F = H^{\text{core}} + G$.

(7) Calculate the transformed Fock matrix $F' = X^H F X$.

(8) Diagonalize F' to obtain C' and ε.

(9) Work out $C = XC'$.

(10) Form a new density matrix D from C using (3.23).

(11) Test to see whether the procedure has converged. This is done by determining whether the new density matrix of step (10) is the same as the previous density matrix to the required accuracy. If the procedure has not converged, go back to step (5) and use the new density matrix.

(12) If the procedure has converged, use the resultant solution, represented by the matrices C, D, F, ... , to calculate the quantities of interest.

The *ab initio* calculation that is being done is completely specified by the choice of basis set and the coordinates of the nuclei. It is important to note that the choice of a basis set is more of an art than a science.

The simplest initial guess at D is the zero matrix, and this is equivalent to approximating F as H^{core} and neglecting all electron-electron interactions in the first iteration. However the SCF procedure will often not converge with this initial guess. So **semi-empirical** calculations are often used to get the initial guess at the wave function. Semi-empirical methods are looked at in section 4.1.

The major time consuming parts of the iteration procedure are in steps (2) and (5), where the two-electron integrals are calculated, and these and the density matrix are put into G. As we will discuss later the time it takes to calculate $(\mu\nu|\lambda\sigma)$ will depend on what type of functions are used. The matrix operations in steps (6)–(10) are not time consuming relative to the formation of G, as long as an efficient diagonalization procedure is used. However as we shall see in chapter 5, it is the diagonalization step that is a bottleneck when the SCF procedure is parallelized.

The procedure will not always converge and even if it does it can be slow. So a lot of techniques have been suggested for ensuring or accelerating convergence. A technique for ensuring convergence, level-shifting, is described in section 5.1 and, a technique for accelerating convergence, direct inversion in the iterative space, is described in section 5.2.

A stopping criterion is needed. We can look at the total electronic energy at each iteration and require that successive values differ by no more than δ. For most purposes $\delta = 10^{-6}$ *Hartrees* will do according to [42]. Alternatively we can require that the standard deviation of successive density matrix elements, which is

$$\left[N_{\text{BF}}^{-2}\sum_{\mu}\sum_{\nu}\left[D_{\mu\nu}^{(i)} - D_{\mu\nu}^{(i-1)}\right]^2\right]^{\frac{1}{2}},$$

to be less than δ. If we use $\delta = 10^{-4}$ then it will usually give an error of less than 10^{-6} *Hartrees* in the energy.

We now look at things that can be calculated from the results of the SCF calculation.

As we have already said in relation to Koopmans' theorem, the values of $-\varepsilon_a$ are a reasonable approximation to the observed IP's, but $-\varepsilon_r$ is usually of little use for EA's. So the eigenvalues ε_a are generally the only ones that are useful numbers as they are.

The total electronic energy is the expectation value $E_0 = \langle\Psi_0|\mathcal{H}_{\text{elec}}|\Psi_0\rangle$ and is given by equation (3.7). Using definition (3.4) for the Fock operator and equation (3.12) for the orbital energy we have

$$f_{aa} = (\psi_a|f|\psi_a) = \varepsilon_a = h_{aa} + \sum_b^{N/2} 2J_{ab} - K_{ab}$$

Therefore by equation (3.7), we can write the energy as,

$$E_0 = \sum_a^{N/2} (h_{aa} + f_{aa}) = \sum_a^{N/2} (h_{aa} + \varepsilon_a)$$

Now if we substitute the basis function expansion (3.14) for the molecular orbitals into this we can show that

$$E_0 = \tfrac{1}{2} \sum_\mu \sum_\nu D_{\mu\nu} \left(H_{\mu\nu}^{\text{core}} + F_{\mu\nu} \right) \tag{3.34}$$

Using this formula the energy can be evaluated from quantities that are available at any stage of the iteration procedure.

If E_0 is calculated from the same D that was used in forming F, then E_0 will be an upper bound to the true energy at any stage of the procedure. When it converges it will usually do so monotonically. By looking back at equation (2.10) we can see that the energy will converge quadratically with respect to the convergence of the wave function.

If we add the classical nuclear-nuclear repulsion term to the electronic energy we will have the total energy E_{tot},

$$E_{\text{tot}} = E_0 + \sum_A \sum_{B>A} \frac{Z_A Z_B}{R_{AB}}$$

It is a function of the nuclear coordinates $\{\mathbf{R}_A\}$. Commonly this is the quantity of most interest, particularly in structure determinations. If the calculation is repeated for different nuclear coordinates then the **potential energy surface** can be explored for nuclear motion. This surface is shown schematically in figure 3.2. It is common to calculate the **equilibrium geometry of a molecule** which involves finding the set $\{\mathbf{R}_A\}$ which minimizes the total energy. The dip in figure 3.2 represents this. This can be done for any collection of point charges and the problem can get quite complicated. If we use a set of nuclear charges that represent more than one molecule (a "**super-molecule**") then **intermolecular forces** can be explored, and the interaction energy can be examined. This will be defined in subsection 3.3.4.

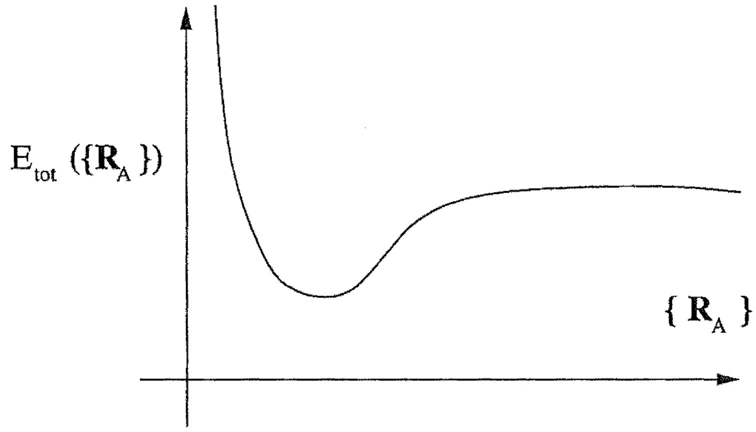


Figure 3.2: Schematic illustration of a potential surface.

The charge density

$$\rho(\mathbf{r}) = \sum_{\mu}\sum_{\nu} D_{\mu\nu}\phi_{\mu}(\mathbf{r})\phi_{\nu}^{*}(\mathbf{r})$$

which represents the probability of finding an electron in various regions, is commonly pictured by contour maps for various planes drawn through the molecule.

There is no unique definition of the number of electrons associated with a given atom or nucleus in a molecule. Since

$$N = 2\sum_{a=1}^{N/2} \int |\psi_a(\mathbf{r})|^2 d\mathbf{r} \tag{3.35}$$

divides the total number of electrons into two electrons per molecular orbital, by substituting the basis expansion of ψ_a into (3.35) we get,

$$N = \sum_{\mu}\sum_{\nu} D_{\mu\nu}S_{\mu\nu} = \sum_{\mu}(DS)_{\mu\mu} = \text{trace}\,(DS)$$

and it is possible to interpret $(DS)_{\mu\mu}$ as the number of electrons to be associated with ϕ_μ. This is a **Mulliken population analysis**. Assuming the basis functions are centred on atomic nuclei the net charge associated with an atom is given by

$$q_A = Z_A - \sum_{\mu \in A} (DS)_{\mu\mu}$$

where Z_A is the charge of nucleus A and the summation index indicates that we only sum over basis functions centred on A. There are many other population analysis schemes. Care has to be taken when looking at their physical significance. They are useful when comparing different molecules using the same type of basis set for each molecule.

Other properties of molecules that can be evaluated from a molecular wave function include the dipole moment, quadrupole moment, field gradient at a nucleus, and diamagnetic susceptibility.

3.3 Types of Basis Functions

In this section general types of basis functions are discussed, and then we look at some specific types of functions that get used. The material is mainly taken from [16] and [42]. Also see chapter one of [35] which is by Dunning and Hay, and chapter one of the 1990 volume of [22] which is by Feller and Davidson.

It is important to remember that no calculated wave function can be better than the basis set from which it is constructed.

3.3.1 Double-Zeta Basis Sets

In qualitative molecular-orbital theory a **linear combination of atomic orbitals** (LCAO) approximation gets used. The molecular orbitals are written as a linear combination of atomic orbitals centred on the atoms in the molecule. This does not have to be done, but most of the calculations that have been performed have done this.

The simplest type of basis set is a **minimal basis set**. Only functions corresponding to the orbitals that would be considered in elementary valence theory are included. This often gives a wave function and energy that are not close to the Hartree-Fock limit. The accuracy can be significantly improved if the number of functions used is doubled.

Consider methane CH_4. With a minimal basis we would have 1s, 2s and 2p functions on carbon, and a 1s function on each of the hydrogen atoms. If we double each of the types of functions so we use two 1s, two 2s and two 2p functions for C, and two 1s functions for each H atom the accuracy is improved. This sort of basis set is called a **double-zeta** (DZ) basis.

Before we can go any further we need some more chemistry, and this is an appropriate time to review some atomic orbital theory.

3.3.2 Quantum Numbers and Atomic Orbitals

An atomic orbital is a region of space where there is a high probability of finding an electron. We look at quantum numbers in relation to atomic orbitals.

The **principal quantum number** n describes the main energy level an electron occupies. The **subsidiary quantum number** is l and it designates the shape of the region of space the electron occupies. It can take the following values

$$\begin{aligned} l &= 0, 1, 2, 3, \ldots, (n-1) \\ &= \text{s, p, d, f, } \ldots \end{aligned}$$

So l designates a sublevel or a kind of atomic orbital. The letter notation shown above gets used. The **magnetic quantum number** m gives the spatial orientation of an atomic orbital. It can take any value below,

$$m = -l,\ -(l-1),\ \ldots,\ -1,\ 0,\ 1,\ \ldots,\ l-1,\ l$$

The fourth quantum number is the **spin quantum number** and it is either $+\frac{1}{2}$ or $-\frac{1}{2}$. This corresponds to spin up and spin down. It gives the spin of the electron and the orientation of the magnetic field produced by this spin. Each main energy level can contain $2n^2$ electrons.

We can represent all the atomic orbitals of an atom with a diffuse cloud of electrons. The 2p orbital is shown in figure 3.3. The "lobe" is actually more diffuse than shown and there is approximately a 90% probability of finding an electron within the surface. The nucleus is in the middle and there is a zero probability of finding an electron there.

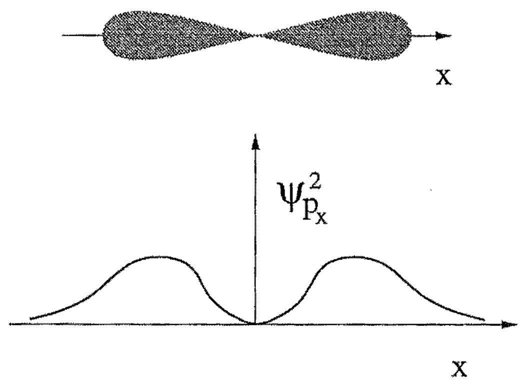


Figure 3.3: The shape of a 2p orbital.

Consider the case $n=2$. We have $l=0$ or $l=1$, so there is an s sublevel and a p sublevel. For the s sublevel $m=0$ is the only possibility. So this sublevel contains two electrons with spin $+\frac{1}{2}$ and spin $-\frac{1}{2}$. Now for the p sublevel m can take the values -1, 0, 1. This gives the spatial orientation of the orbitals, and we have three orbitals p_x, p_y and p_z. These are shown in figure 3.4. Each of them contains two electrons.

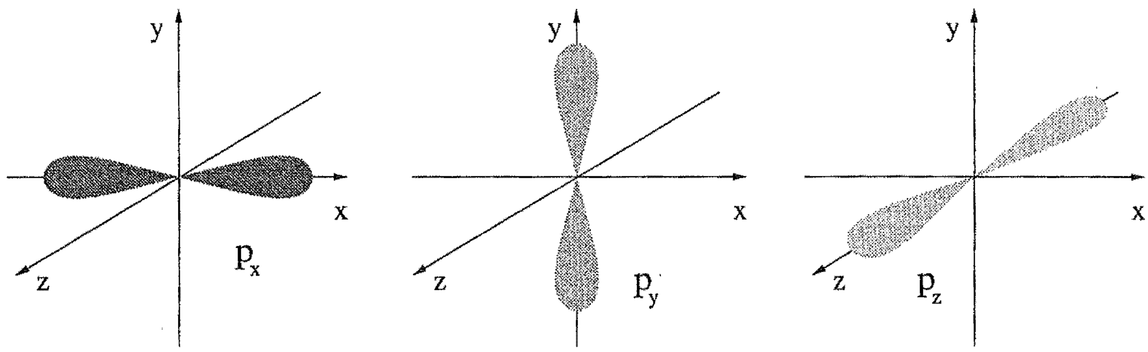


Figure 3.4: The three atomic p orbitals shown separately.

Successive energy levels are at increasingly greater distances from the nucleus. Each energy level has one s sublevel consisting of one s atomic orbital. Each s orbital is spherically symmetric with respect to the nucleus. Each p sublevel consists of a set of three p atomic orbitals, and they look like three perpendicular equal-arm dumbbells or lobes.

Higher energy levels have more complicated shapes and these are described in [45, pages 111–113].

3.3.3 Double-Zeta-Plus-Polarization Basis Sets

By using atomic functions for molecular orbitals the fact that when a bond is formed one atom will distort the atomic orbitals of adjacent atoms gets ignored. If orbitals with higher values of the l quantum number are included then this can be taken into account.

So for hydrogen the distortion of the 1s orbital can be described by including functions of p character. Such functions get called **polarization functions**. Now back to methane. If we add a set of p functions for each H atom and a set of d functions for C to the DZ basis we obtain a **double-zeta-plus-polarization** (DZP) basis. This sort of basis set can give a reasonable description of a molecule in the ground or low excited states.

3.3.4 Basis Set Superposition Error

A problem can come up when we calculate the **interaction energy** for a weakly bound system. If two systems X and Y are separated by a distance r then this energy is given by

$$\Delta E_{XY}(r) \equiv E_{XY}(r) - E_X - E_Y \tag{3.36}$$

Here E_{XY} is the energy of the super-molecule XY, and E_X and E_Y are the energies of the separated systems.

If we use basis sets S_X and S_Y for X and Y respectively, then the basis set for XY will be $S_X \cup S_Y$. The use of a basis set with a finite number of functions leads to a **basis set truncation error** ϵ_i for species i. In general

$$\epsilon_{XY} - \epsilon_X - \epsilon_Y \neq 0$$

In the calculation for super-molecule XY we get a better description of X than in the calculation of X alone because we use the bigger basis set $S_X \cup S_Y$. The inclusion of the S_Y basis functions results in a *non-physical* lowering of the energy of XY. Similarly we also get a better description of Y in XY. This energy lowering is known as **basis set superposition error.**

An approximate way of taking account of this involves calculating the energies E_X and E_Y of the systems X and Y using the full basis set $S_X \cup S_Y$.

This error does not come up in any later discussions but it is interesting to note one of the causes of error.

3.3.5 Slater-Type and Gaussian-Type Functions

We now look at the nature of the functions in some basis sets.

Slater-type functions, or **Slater-type orbitals** (STO's), are of the form

$$\phi_{nlm}^{\text{STO}}(\zeta, r) = r^{n-1} \exp(-\zeta r) Y_{lm}(\theta, \lambda) \tag{3.37}$$

where n, l and m are the usual atomic quantum numbers, $Y_{lm}(\theta, \lambda)$ is a spherically symmetric harmonic function, and ζ is the Slater orbital exponent parameter. It should be noted that ζ is a parameter in the actual SCF calculation. These sorts of functions are hydrogen-like because they are used to describe **hydrogen-like atoms**, which are a nucleus with only one electron ie. H, He^+, Li^{2+}, ...

The orbital exponents are chosen in various ways; by using a set of rules, by performing an atomic SCF calculation, or by optimizing their values in a molecular SCF calculation.

A major disadvantage with Slater-type functions is that the two-electron integrals $(pq|rs)$, which have the four functions on three or four different centres are very difficult to evaluate accurately enough, and are therefore very time consuming.

It was suggested that Cartesian Gaussian functions be used to make the integral calculations faster. These are of the form

$$\phi_{lmn}^{\text{GTO}}(\alpha, r) = x^l\, y^m\, z^n \exp(-\alpha\, r^2) \tag{3.38}$$

(The indices l, m and n are not the usual atomic quantum numbers this time.) Here α is the Gaussian orbital exponent. Also r is not $|\mathbf{r}|$ here but $|\mathbf{r} - \mathbf{R}_A|$ where A is the nucleus that the function is centred on. These get called **Gaussian-type orbitals** (GTO's).

We can define $L = l + m + n$, and refer to $L = 0$ functions as "s" functions, $L = 1$ functions as "p" functions, etc.

Strictly speaking both the STO's and GTO's should be normalized. This is not necessary for the present discussion. A simple STO and a simple GTO are shown in figures 3.5 and 3.6 respectively.

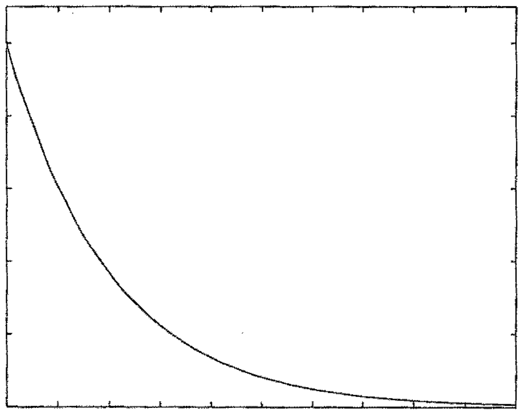


Figure 3.5: A Slater-type orbital.

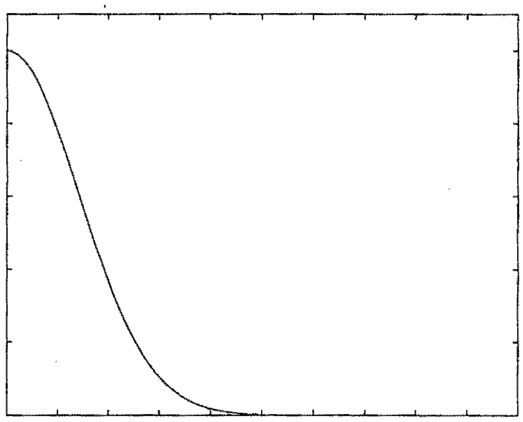


Figure 3.6: A Gaussian-type orbital.

In contrast to Slater functions where 1s, 2s, 3s, ... functions depend on r as $e^{-\zeta r}, re^{-\zeta r}, r^2 e^{-\zeta r}, \ldots$, all "s" Gaussian functions are taken to decay as $e^{-\alpha r^2}$. So only Gaussian "1s" functions are used.

In the two-electron integral $(\mu\nu|\lambda\sigma)$ the product $\phi_\mu \phi_\nu$, with ϕ_μ and ϕ_ν on different centres can be replaced by a single Gaussian function at a centre between μ and ν on a line joining their two centres. Consider the product of two Gaussian s–functions

$$\exp(-\alpha_\mu r_\mu^2)\exp(-\alpha_\nu r_\nu^2) = \exp\left(-\frac{\alpha_\mu \alpha_\nu}{\alpha_\mu + \alpha_\nu} r_{\mu\nu}^2\right)\exp(-\alpha r_C^2)$$

where $r_{\mu\nu}$ is the distance between the centres of ϕ_μ and ϕ_ν and is given by

$$r_{\mu\nu}^2 = (\mu_x - \nu_x)^2 + (\mu_y - \nu_y)^2 + (\mu_z - \nu_z)^2$$

where it is given in terms of the coordinates of the two centres. The orbital exponent is

$$\alpha = \alpha_\mu + \alpha_\nu$$

and r_C is the distance from centre C which has coordinates

$$C_i = \frac{\alpha_\mu \mu_i + \alpha_\nu \nu_i}{\alpha_\mu + \alpha_\nu} \qquad\qquad i = x, y, z$$

As we might expect, this advantage is partly offset by the fact that GTO's give a less effective description of atomic wave functions. For example, an exponential 1s function has a cusp at the nucleus whereas there is no cusp in a Gaussian function. Even if the orbital exponent is chosen so that the functions match up quite well in some intermediate region, the Gaussian function will also deviate in the long range due to the differing decay rates. Gaussian functions decay much more rapidly. It can be shown that molecular orbitals decay as $\exp(-\zeta r)$ at large distances, which is how Slater functions decay.

3.3.6 Contracted Gaussian Basis Sets

If we want to get an accuracy with Gaussian functions that is comparable to what we can get with Slater-type exponential functions we obviously need to use a bigger basis set. We might need three times as many Gaussian functions as Slater functions. This makes things slower since the time taken to construct the Fock matrix depends on N_{BF}^4, where N_{BF} is the number of basis functions.

This problem has been solved by using **contracted Gaussian functions**, where some of the primitive Gaussian functions are grouped together in fixed linear combinations. We can do this by finding linear combinations of Gaussian functions that approximate a Slater-type orbital. Consider an expansion in terms of three Gaussian functions

$$\phi^{\text{CGF}} = d_1 \exp(-\alpha_1 r^2) + d_2 \exp(-\alpha_2 r^2) + d_3 \exp(-\alpha_3 r^2) \tag{3.39}$$

The coefficients d_i and orbital exponents α_i are chosen to give the least-squares fit to a Slater-type function. This is known as a STO-3G orbital, and the combination is left the same right through a SCF calculation. More generally STO-LG represents the expansion of a Slater-type function in terms of L Gaussian functions. (The notation STO-NG is usually used but N is the number of electrons here.) For minimal basis sets this is reasonable but otherwise it is too inflexible. In [42, pages 159–180] SCF calculations on H_2 and $He\,H^+$ are done using a STO-3G basis set.

The contracted Gaussian basis sets are generally derived from sets of primitive Gaussian functions obtained in atomic SCF calculations. The most efficient contraction schemes leave the smallest exponent primitive functions uncontracted and make a single contraction of the largest exponent ones. Therefore the outermost functions are not contracted and the innermost primitives are. This is so the outer functions, which have the biggest amplitude in the inter-atomic regions, can respond to the changes that occur when the molecule forms. The inner functions in a sense describe atomic regions. Sometimes, because different orbitals have the same symmetry characteristics, a function is included in more than one contracted function or left uncontracted.

Generally speaking if we want to get higher accuracy in molecular calculations it is better to include extra functions in the basis set than to try and optimize orbital exponents exhaustively.

The results of SCF calculations on simple poly-atomic molecules are considered in [42, pages 180–190]. Different basis sets are compared.

The reader is referred to the things in the bibliography mentioned at the start of the section for further details on basis functions.

Chapter 4

Configuration Interaction

The Hartree-Fock neglect of electron correlation has two consequences; the results are not accurate, and the wave function will behave incorrectly as the molecule dissociates. We now look at more advanced methods that take more account of electron correlation than the Hartree-Fock SCF procedure. The methods we look at fall under the heading configuration interaction. As the name suggests these methods use (linear) combinations of different configurations to give a better wave function. Before we look at these methods we take a brief and general look at semi-empirical methods.

4.1 Semi-Empirical Methods

The Hartree-Fock SCF procedure we have been looking at is an *ab initio* method. We now compare *ab initio* and semi-empirical methods.

With *ab initio* methods an appropriate model for the molecular wave function is chosen and then the calculations are performed without further approximations or input from experiment. The approximation lies in the choice of the model, and an inappropriate model leads to a result that is inaccurate and might be misleading. So *ab initio* does not necessarily imply accuracy. It is possible to use *ab initio* methods to get accurate quantitative results for small molecules as long as a large enough basis set is used and electron correlation is treated adequately enough. In the rest of this chapter we look at *ab initio* methods that treat electron correlation. If we consider larger molecules the results become a lot more qualitative, and they might give insight into a particular chemical situation rather than numerical results.

As *ab initio* methods are only applicable to a relatively small number of problems chemists have spent a lot of time devising computationally simpler schemes for making electronic structure calculations for a wider variety of systems. A lot of these methods involve the use of experimental information and are called **semi-empirical**. This sort of approach takes into account that it is very hard to get energies that are of **chemical accuracy** which is within 4 $kJ\,mol^{-1}$ of experiment. Semi-empirical methods are derived by taking the exact equations of *ab initio* theory and making approximations. They generally use a simpler Hamiltonian, and incorporate experimental data or parameters that can be adjusted to fit experimental data. They drastically reduce the number of two-electron integrals that need to be calculated.

Stewart, Császár and Pulay [41] show how SCF-type semi-empirical methods can be accelerated. These methods differ from the *ab initio* ones we are looking at mainly in the fact that the matrix diagonalization makes up the main computational cost. The article [41] proposes a method that does not require full diagonalization but still works well.

4.2 The Idea Behind Configuration Interaction

The book by Hirst [16] gives a basic outline of most of the material covered in this section. We begin by looking at the basic idea of configuration interaction. Then different types of multi-configuration self-consistent field and configuration interaction methods are discussed.

In its simplest form **configuration interaction** (CI) is the expansion of a molecular wave function as a sum of Slater determinants $|\Psi_I\rangle$

$$|\Psi\rangle = \sum_I c_I |\Psi_I\rangle \tag{4.1}$$

If the Slater determinants $|\Psi_I\rangle$ are chosen appropriately we can make sure that the wave function behaves correctly when the nuclei are separated. The coefficients c_I come from a variational calculation in which the energy E_0,

$$E_0 = \frac{\langle\Psi|\mathcal{H}_{\text{elec}}|\Psi\rangle}{\langle\Psi|\Psi\rangle}$$

is minimized. Incorrect dissociative behaviour of the Hartree-Fock wave function can be corrected by a small CI calculation. So even a small calculation of this type can allow for non-dynamical correlation which is an important contribution to the correlation energy. To account properly for dynamical correlation an extensive CI calculation needs to be done.

CI is conceptually simple compared to a lot of other methods that get used. With an appropriate choice of Ψ_I CI corrects for the neglect of electron correlation and is possibly the most widely applicable method for the calculation of accurate molecular wave functions. An alternative approach is to use the **generalized valence-bond** (GVB) method which is described in [16, page 41].

The methods we look at in this section are also applicable to open-shell molecules. The details of this case are not too important to the discussion, but it is important to keep in mind that what we are talking about is a lot more general than the material in the previous chapter.

In what get referred to as **configuration interaction methods** the functions $|\Psi_I\rangle$ are fixed and the coefficients c_I are chosen to minimize the energy E_0. These functions need to have the same symmetry and spin as the desired wave function. They are called **configuration state function** (CSF's).

In the **multi-configuration self-consistent field** (MCSCF) method we consider the mixing of several configurations as well as the orbitals of which they are composed. So we optimize both the linear variational coefficients c_I, corresponding to different configurations, *and* the expansion coefficients $C_{\mu i}$, corresponding to the basis functions. This is done simultaneously and makes a MCSCF calculation much more difficult computationally than a conventional SCF calculation.

If we consider the same number of CSF's a MCSCF calculation is much more difficult computationally than one done using a CI method. So in the MCSCF method a more limited number of configurations is considered.

A method is **size consistent** if the energy of a system consisting of two subsystems A and B at infinite separation is the same as the sum of the energies of A and B calculated separately by the same method. Truncation of the configuration list may lead to a wave function that is not size consistent. **Many-body perturbation theory** is a size consistent method, but we will not discuss it here. It is looked at in [42, pages 320–379].

In the next two subsections two different MCSCF optimization schemes are discussed. We will not look at any actual MCSCF algorithms in this chapter or any of the chapters that follow. However it is important to get an idea of what goes into a MCSCF calculation. Also the MCSCF method is mentioned throughout section 5.3.

In what follows the orbital basis is called the **atomic orbital** (AO) basis, and the orthonormal orbitals that define the wave function are the **molecular orbitals** (MO'S).

4.2.1 Complete Active-Space SCF (a MCSCF method)

In 1980 Roos, Taylor and Siegbahn [29] proposed the **complete active-space self-consistent field** (CASSCF) method. It minimizes the element of choice in selecting the configurations. It is discussed by Roos in [19, pages 399–446]. CASSCF attempts to keep as much of the conceptual simplicity of the Hartree-Fock approach as possible.

The molecular orbitals are divided into three sets. The lowest orbitals are doubly occupied in all configurations and are called the **inactive** set. The highest orbitals are the **virtual** orbitals and are unoccupied in all configurations. In between are the **active** orbitals.

The configuration list consists of those configurations, of appropriate symmetry and spin, that can be generated from all possible arrangements of the active electrons among the active orbitals. There is no possibility of overlooking a configuration that could be important. The choice is involved in selecting the set of active orbitals.

We need to generalize the density matrix of equation (3.23) which was for the closed-shell case. We generalize it so that orbital a has n_a electrons in it rather than being restricted to having 2 electrons. So n_a is equal to 0, 1 or 2, and the charge density is

$$\rho(\mathbf{r}) = \sum_a n_a \, |\psi_a(\mathbf{r})|^2 = \sum_{a\mu\nu} n_a \, C_{\mu a} \, C^*_{\nu a} \, \phi_\mu(\mathbf{r}) \, \phi^*_\nu(\mathbf{r}) \tag{4.2}$$

Note we are now summing over all the molecular orbitals rather than just the (doubly) occupied ones. As the ψ_a do not need to be doubly occupied now this definition applies to open-shell wave functions too.

For an N–electron system with wave function $\Psi(\mathbf{x}_1, \ldots, \mathbf{x}_N)$ the probability of finding any electron in the volume $d\mathbf{r}_1$, irrespective of spin, is

$$P_1(\mathbf{r}_1; \mathbf{r}_1') = N \int \Psi(\mathbf{r}_1, \omega_1, \mathbf{x}_2, \ldots, \mathbf{x}_N) \Psi^*(\mathbf{r}_1', \omega_1, \mathbf{x}_2, \ldots, \mathbf{x}_N) d\omega_1 d\mathbf{x}_2 \cdots d\mathbf{x}_N$$

where ω is the spin variable and the integral is with respect to the spin variable of all the electrons. Recall that $\mathbf{x}_i = \{\mathbf{r}_i, \omega_i\}$. The variable $\mathbf{r}_1'$ is introduced in Ψ^* so that the effect of a one-electron operation on $P_1(\mathbf{r}_1; \mathbf{r}_1')$ can be limited to the contribution from Ψ. As far as we have been concerned so far $\mathbf{r}_1 = \mathbf{r}_1'$. The factor of N is a normalization constant. The *function* $P_1(\mathbf{r}_1; \mathbf{r}_1')$ is the **spinless first-order density matrix**. It is referred to as a matrix even though strictly speaking it is a function of two continuous variables. For a closed-shell SCF wave function Ψ is a single Slater determinant and

$$P_1(\mathbf{r}_1; \mathbf{r}_1') = \sum_{\mu\nu} D_{\mu\nu} \phi_\mu(\mathbf{r}_1) \phi_\nu(\mathbf{r}_1')$$

where $D = 2CC^T$. If $\mathbf{r}_1 = \mathbf{r}_1' = \mathbf{r}$ we have (3.22).

Now the **spinless second-order density matrix** gives the probability of there being an electron in $d\mathbf{r}_1$ and another electron in $d\mathbf{r}_2$ irrespective of spin. It is given by

$$\begin{aligned} P_2(\mathbf{r}_1, \mathbf{r}_2; \mathbf{r}_1', \mathbf{r}_2') &= N(N-1) \\ \int \Psi(\mathbf{r}_1, \omega_1, \mathbf{r}_2, \omega_2, \mathbf{x}_3, \ldots, \mathbf{x}_N) & \Psi^*(\mathbf{r}_1', \omega_1, \mathbf{r}_2', \omega_2, \mathbf{x}_3, \ldots, \mathbf{x}_N) d\omega_1 d\omega_2 d\mathbf{x}_3 \cdots d\mathbf{x}_N \end{aligned}$$

The MCSCF wave function Ψ is expressed in terms of a linear combination of Slater determinants written in terms of molecular orbitals ψ_i. The spinless first-order reduced density matrix can be written as

$$P_1(\mathbf{r}_1; \mathbf{r}_1') = \sum_{ij} \gamma_{ij} \psi_i(\mathbf{r}_1) \psi_j(\mathbf{r}_1')$$

The γ_{ij} depend on the CI coefficients c_I and on the structures of the Slater determinants comprising Ψ. Similarly the second-order density matrix can be expanded in terms of molecular orbitals too with coefficients Γ_{ijkl}. The coefficients γ_{ij} and Γ_{ijkl} are important because the energy for a MCSCF wave function is

$$E_{\text{MCSCF}} = \sum_{ij} h_{ij}\, \gamma_{ij} + \tfrac{1}{2} \sum_{ijkl} g_{ijkl}\, \Gamma_{ijkl}$$

The h_{ij} and g_{ijkl} are the one-electron and two-electron integrals with respect to the molecular orbitals ψ_i from which the CSF's Ψ_I are built up. So

$$h_{ij} = \int \psi_i(\mathbf{r}_1) h(\mathbf{r}_1) \psi_j(\mathbf{r}_1) d\mathbf{r}_1$$

and

$$g_{ijkl} = \int \psi_i(\mathbf{r}_1)\psi_j(\mathbf{r}_1) r_{12}^{-1} \psi_k(\mathbf{r}_2)\psi_l(\mathbf{r}_2) d\mathbf{r}_1 d\mathbf{r}_2 \tag{4.3}$$

These were first defined in subsection 3.1.2, with different notation for g_{ijkl}. We can write γ_{ij} and Γ_{ijkl} in terms of the c_I's by putting

$$\gamma_{ij} = \sum_I \sum_J c_I c_J \gamma_{ij}^{IJ}$$

and

$$\Gamma_{ijkl} = \tfrac{1}{2} \sum_I \sum_J c_I c_J \left(\Gamma_{ijkl}^{IJ} + \Gamma_{jikl}^{IJ} \right)$$

The γ_{ij}^{IJ} and Γ_{ijkl}^{IJ} are **electron spin coupling coefficients**. They depend on the structure of the wave function and can be evaluated straightforwardly. The energy is optimized with respect to the coefficients c_I that appear in γ_{ij} and Γ_{ijkl}, and with respect to the coefficients $C_{\mu i}$ which are in h_{ij} and g_{ijkl}.

The optimization of the molecular orbital coefficients can be expressed using a unitary matrix U. Suppose $\{\psi_i\}$ is an initial set of molecular orbitals and $\{\psi_i'\}$ is an improved set, putting them into vectors gives

$$\boldsymbol{\psi}' = \boldsymbol{\psi}\, U$$

A unitary matrix can be represented in exponential form, so we can put

$$U = \exp R = I + R + \tfrac{1}{2} R^2 + \cdots$$

where R is antisymmetric. The entries in R form an independent set of parameters that can be used to describe the orbital rotations. We will look more at how this sort of parameterization works later on page 84.

The optimization of the coefficients c_I can be described in terms of a unitary matrix too. We can express it as $\exp S$ where the elements of S describe the mixing of the wave function Ψ.

Thus an improved wave function Ψ' can be expressed as

$$\Psi' = \Psi \exp S \exp R$$

where Ψ is the starting function. The elements of R and S are the variational parameters and they must be optimized to get the minimum energy.

Using a quadratic Taylor expansion we can express the energy as

$$
\begin{aligned}
E^{(2)}(\mathbf{r},\mathbf{s}) &= E^{(0)} + [\mathbf{s}^T\,\mathbf{r}^T]\begin{bmatrix} \mathbf{g}^{(c)} \\ \mathbf{g}^{(o)} \end{bmatrix} \\
&+ \tfrac{1}{2}[\mathbf{s}^T\,\mathbf{r}^T]\begin{bmatrix} B^{(cc)} & B^{(co)} \\ B^{(co)T} & B^{(oo)} \end{bmatrix}\begin{bmatrix} \mathbf{r} \\ \mathbf{s} \end{bmatrix}
\end{aligned}
$$

where $E^{(0)}$ is the energy of the original wave function Ψ. The vectors $\mathbf{r}$ and $\mathbf{s}$ contain all the elements of the matrices R and S respectively. The vectors $\mathbf{g}^{(c)}$ and $\mathbf{g}^{(o)}$ are the gradients of the energy with respect to the CI coefficients and the orbital coefficients respectively. The matrix B is a Hessian with the superscripts defined in the obvious way. The energy will be stationary when

$$
\begin{bmatrix} \mathbf{g}^{(c)} \\ \mathbf{g}^{(o)} \end{bmatrix} + \frac{1}{2}\begin{bmatrix} B^{(cc)} & B^{(co)} \\ B^{(co)T} & B^{(oo)} \end{bmatrix}\begin{bmatrix} \mathbf{r} \\ \mathbf{s} \end{bmatrix} = 0
$$

In principle the Newton-Raphson method can be used for optimization and when this is used the improved set of variation parameters is given by $-2B^{-1}\mathbf{g}$. If we start far from the minimum, the energy will not be accurately represented by a quadratic and the Hessian might have negative or very small eigenvalues. A shifting technique can be used to make the Hessian positive definite so that satisfactory convergence is achieved.

4.2.2 Super-CI (a MCSCF method)

An alternative method of MCSCF optimization is the **super configuration interaction** (super-CI). It was first proposed by Grein and Chang [10] back in 1971 and is based on the generalized Brillouin theorem.

Suppose $|\Psi_0\rangle$ is the wave function from a closed-shell SCF calculation, and $|\Psi_a^r\rangle$ is the wave function obtained from $|\Psi_0\rangle$ by exciting an electron from one of the occupied orbitals a to one of the virtual orbitals r. Then by the Brillouin theorem of subsection 2.2.6

$$\langle\Psi_0|\mathcal{H}_{\text{elec}}|\Psi_a^r\rangle = 0$$

This was generalized to the open-shell case and MCSCF wave functions in 1968 [21]. We have expressed the MCSCF wave function as

$$|\Psi\rangle = \sum_I c_I |\Psi_I\rangle \tag{4.4}$$

Construct a new wave function $|\Psi(i\to j)\rangle$ by swapping an electron from orbital i to orbital j *and* from orbital j to orbital i as follows

$$|\Psi(i\to j)\rangle \equiv \sum_I c_I \left[\, |\Psi_I{}_i^j\rangle - |\Psi_I{}_j^i\rangle \,\right] \tag{4.5}$$

If any of the excitations are not possible (ie. ψ_i or ψ_j is not occupied in $|\Psi\rangle$), or are prohibited by the Pauli principle, then we set the excited wave function to zero. So if $|\Psi_I{}_i^j\rangle$ is not possible we set $|\Psi_I{}_i^j\rangle = 0$ in (4.5). We can now state the theorem. It will not be proved here.

Generalized Brillouin Theorem: *If $|\Psi(i\to j)\rangle$ is constructed according to equation (4.5) then*

$$\langle\Psi|\mathcal{H}_{\text{elec}}|\Psi(i\to j)\rangle = 0 \tag{4.6}$$

Now that we have this result we can look at the method. A **super-CI wave function** is constructed as

$$|\Psi_{\text{SCI}}\rangle = |\Psi\rangle + \sum_{i<j} x_{ij}\, |\Psi(i\to j)\rangle$$

The linear variational problem is solved to give the coefficients x_{ij}. This is a CI problem, with CI coefficients $\{\, 1,\ x_{ij} \ \ i<j \,\}$, and hence the name super-CI. These coefficients are then used to define the new molecular orbitals ψ_i' using the formula

$$\psi_i' = \psi_i + \sum_{j\neq i} x_{ij}\psi_j$$

with $x_{ij} = -x_{ji}$. Using the new orbitals the CI coefficients in equation (4.4) are calculated giving an improved MCSCF wave function. The process is repeated until the x_{ij} are zero and the generalized Brillouin theorem is satisfied. When the x_{ij} are zero the super-CI wave function is not improved at all by adding any of the $|\Psi(i\to j)\rangle$ to it and the molecular orbitals have converged. Once this has happened (4.6) is satisfied. This is clearer once we look at the actual equations that are used to solve the CI problem. What happens is we get the eigenvector of the Hamiltonian matrix equal to $\mathbf{e}_1$ (as the CI coefficient of $|\Psi\rangle$ is the only nonzero one), and the energy (eigenvalue) is equal to the (1,1)-component of the Hamiltonian matrix. The off-diagonal elements in the Hamiltonian are all zero which is what the generalized Brillouin theorem says.

An alternative method of getting the ψ_i''s is to calculate the reduced first-order density matrix for the wave function $|\Psi_{\text{SCI}}\rangle$ in the atomic orbital basis. Diagonalization of this matrix yields a set of orbitals called the **natural orbitals** $\bar{\psi}_i$, which are such that we can write

$$P_i(\mathbf{r}_1;\mathbf{r}_1') = \sum_i n_i \bar{\psi}_i(\mathbf{r}_1)\bar{\psi}_i(\mathbf{r}_1')$$

where the occupation numbers n_i are the eigenvalues of the density matrix. The natural orbitals with the highest occupation numbers are used as the improved molecular orbitals ψ_i', so we are looking for the largest eigenvalues. With a restricted closed-shell Hartree-Fock wave function the natural orbitals are the molecular orbitals.

One effective strategy for optimizing a MCSCF wave function is to start with a super-CI method and once convergence has kicked in change to the second-order Newton-Raphson method.

4.2.3 The Full-CI Method

In this method we just optimize the coefficients c_I in

$$|\Psi\rangle = \sum_I c_I |\Psi_I\rangle \tag{4.7}$$

The $|\Psi_I\rangle$ are not necessarily Slater determinants but for now we will assume they are.

The best molecular electronic wave function that can be calculated with a given basis set is obtained by performing a **full configuration interaction** (full-CI) calculation. We consider all possible excitations of electrons from the occupied orbitals of the SCF or RHF wave function to virtual orbitals. (Recall that RHF is the restricted Hartree-Fock method and is used for the calculation of open-shell SCF wave functions.) Each of these excitations will give a Slater determinant but it may not have a well defined symmetry or spin angular momentum.

For a moment assume we can keep all the Slater determinants. Then the form of the full-CI wave function is

$$|\Psi\rangle = c_0|\Psi_0\rangle + \sum_{ra} c_a^r|\Psi_a^r\rangle + \sum_{\substack{a<b\\ r<s}} c_{ab}^{rs}|\Psi_{ab}^{rs}\rangle + \sum_{\substack{a<b<c\\ r<s<t}} c_{abc}^{rst}|\Psi_{abc}^{rst}\rangle + \dots \tag{4.8}$$

which we can write symbolically as

$$|\Psi\rangle = c_0|\Psi_0\rangle + c_S|S\rangle + c_D|D\rangle + c_T|T\rangle + c_Q|Q\rangle + \dots$$

where $|S\rangle$ represents terms involving single excitations, $|D\rangle$ represents terms involving double excitations and so on. It is important to remember that these sums are finite. Recall that N is the number of electrons and N_{BF} is the number of basis functions, which means there are $\binom{2N_{\text{BF}}}{N}$ determinants involved in such a calculation.

We actually need to form linear combinations which have the same symmetry and spin as the SCF wave function Ψ_0. As we have already mentioned these functions Ψ_I are called configuration state functions (CSF's). The coefficients c_I are chosen so that the energy E_0 is a minimum and this calculation will be looked at in subsection 4.2.6.

For most molecules with a reasonably good basis set a full-CI wave function consists of such a large number of CSF's Ψ_I that a full CI calculation is prohibitively large. So such a calculation is only possible for relatively small systems. In general we have to truncate the configuration list. We should start with a set of MO's that give a reasonable description of the system. All forms of truncated CI deteriorate as the number of electrons increases. For most molecules the orbitals obtained in a RHF calculation will be a suitable zeroth-order wave function if we are near the equilibrium configuration. In cases of near degeneracy or for problems where the geometry is not near the equilibrium configuration orbitals from a GVB, MCSCF or CASSCF calculation need to be used.

4.2.4 Singly and Doubly Excited CI

We need to look at criteria for truncating the configuration list. For now assume the SCF or RHF wave function $|\Psi_0\rangle$ is a reasonable approximation to the exact wave function $|\Phi\rangle$. It can be shown that

$$\langle\Psi_i|\mathcal{H}_{\text{elec}}|\Psi_j\rangle = 0$$

where $|\Psi_i\rangle$ and $|\Psi_j\rangle$ are Slater determinants that differ by more than two spin orbitals. So there will be no *direct interaction* between $|\Psi_0\rangle$ and a CSF for which more than two electrons have been promoted from orbitals occupied in $|\Psi_0\rangle$. We can have $|\Psi_a^r\rangle$ interacting with $|\Psi_{abc}^{rst}\rangle$ say. Limitation of the configuration list to single and double excitations from $|\Psi_0\rangle$ therefore seems like a reasonable first approximation. This is known as **singly and doubly excited configuration interaction** (SDCI). If a system has only two electrons then SDCI obviously corresponds to full-CI.

The Hartree-Fock approximation takes account of a large part of the correlation of electrons with parallel spins through the antisymmetry condition. The main part of the remaining error is due to the correlation of electron pairs having opposite spins. Electrons correlation to a good approximation is described by pair interactions [35, page 278], so the dominant part of the correlation energy is obtained with a CI wave function which contains excitations in each pair of electrons. Single excitations are also important for other reasons.

In general, a wave function describing dynamical correlation effects should contain the near-degenerate configurations and single and double excitations with respect to all of them. This is the minimum requirement.

In the closed-shell case, due to Brillouin's theorem,

$$\langle\Psi_0|\mathcal{H}_{\text{elec}}|\Psi_I^{(1)}\rangle = 0$$

where $\Psi_I^{(1)}$ is a CSF with only one electron excited. So in the closed-shell case single excitations do not interact with $|\Psi_0\rangle$ and therefore will have a negligible effect on the energy. They can however be important in the calculation of molecular properties for closed-shell molecules. For open-shell molecules single excitations are very important and must be included in the configuration list. In the closed-shell case perturbation theory shows that including double excitations gives a first-order correction to the wave function, and makes the energy correct to the third-order. The second-order correction to the wave function is given by single, triple and quadruple excitations.

4.2.5 Multi-Reference CI

A configuration list with single and double excitations with respect to $|\Psi_0\rangle$ is only adequate if $|\Psi_0\rangle$ is the dominant configuration in the CI wave function (so it has the largest CI coefficient). If we want to get the correct dissociative behaviour we usually need to include configurations other than those that come from the Hartree-Fock wave function $|\Psi_0\rangle$. Also the excited states with the same spin and symmetry as $|\Psi_0\rangle$ will be described by something other than excitations with respect to $|\Psi_0\rangle$. So a configuration list with only single and double excitations with respect to $|\Psi_0\rangle$ will not be good enough for these problems. We need to choose a set of reference configurations $\{|\Psi_K^{(R)}\rangle\}$ that includes all the important configurations for the states we are interested in. So we need to include things for the equilibrium region and the asymptotic regions. The minimum we should have is all the configurations with coefficient c_I larger than some specified number, say 0.1, in the final (normalized) CI wave function. So we want the most dominant configurations. We then get the configuration list by taking single and double excitations with respect to each of the reference configurations $|\Psi_K^{(R)}\rangle$. This is **multi-reference configuration interaction** (MRCI).

We now turn to the actual calculation of the CI coefficients.

4.2.6 Conventional and Direct CI

There are a couple of approaches for evaluating the coefficients once we have decided which configurations we will use.

In the **conventional configuration interaction method** we optimize the CI coefficients by *setting up* the **secular equations** which are

$$H\mathbf{c} = E\,S\mathbf{c} \tag{4.9}$$

where H and S are matrices given by

$$H_{IJ} \equiv \langle\Psi_I|\mathcal{H}_{\text{elec}}|\Psi_J\rangle$$

$$S_{IJ} \equiv \langle\Psi_I|\Psi_J\rangle$$

The energy is given by E and $\mathbf{c}$ is a column vector containing the coefficients c_I. The matrix H is a Hamiltonian matrix and it gives the interactions between the CSF's.

Assume the CSF's $|\Psi_I\rangle$ are linear combinations of Slater determinants that are made from orthonormal MO's ψ_i. So the functions $|\Psi_I\rangle$ constructed from different orbital occupancies are orthonormal too. Non-orthonormality between subsets of CSF's can be dealt with by a relatively small Schmidt orthogonalization so that S is equal to the identity. The secular equations are then

$$H\mathbf{c} = E\,\mathbf{c} \tag{4.10}$$

Suppose the total number of CSF's is n. After we get the set $\{c_I^{(1)}\}$ for which the energy is a minimum there will be $n-1$ independent sets of coefficients $\{c_I^{(k)}\}$ (eigenvectors) with eigenvalues E_k. If

$$E_1 \leq E_2 \leq E_3 \leq \cdots \leq E_n$$

then each E_k is an upper bound to the energy of a state of the molecule. So the first m excited states of a given symmetry and spin can be obtained by calculating the first m eigenvalues in (4.10). If the CSF's have been chosen for the ground state these energy values might not be very good approximations.

In the conventional CI method we construct H and calculate the required eigenvalues and eigenvectors. But this limits the number of configurations that can be handled because of storage. So other methods have been developed that can handle much larger numbers of configurations. These methods avoid the explicit construction of the Hamiltonian matrix H and are known as **direct configuration interaction** methods. (It is important to note the meaning of direct here. It comes from the fact that the elements of H are constructed *directly* from the MO's. This should not be confused with the meaning of direct when used to describe eigenvalue methods. Eigenvalue methods are described as direct or iterative. However direct-CI is used with iterative eigenvalue methods. In the next chapter we will consider direct-SCF methods.)

We now briefly look at a problem which is common to conventional-CI, direct-CI and MCSCF methods. The configurations $|\Psi_I\rangle$ are expressed in terms of Slater determinants that are built up from MO's ψ_i. If we have N_{MO} molecular orbitals the number of two-electron integrals g_{ijkl} will be $O(N_{\text{MO}}^4)$. Substituting the MO's ψ_i in terms of N_{BF} basis function ϕ_μ into equation (4.3) results in a four-fold summation over the N_{BF} basis functions. So the work involved in the generation of the g_{ijkl} will be $O(N_{\text{MO}}^4 N_{\text{BF}}^4)$. In a CI calculation usually relatively few of the virtual orbitals are discarded so the four-index transformation is $O(N_{\text{BF}}^8)$. However this can be reduced to $O(N_{\text{BF}}^5)$ by splitting it up into a sequence of four partial summations. These are given in [16, page 54]. This is still a lot of work when N_{BF} is large.

We will look at CI more in chapter 6 where Davidson's method is looked at. This method is used to solve the eigenvalue problem given by (4.10).

4.2.7 Summary of the Methods

Table 4.1 summarises the chemistry methods we have looked at so far and it includes the ones that will be looked at in the next two chapters. The table is divided into three parts for SCF, MCSCF and CI methods.

method	*section(s)*
Hartree-Fock SCF	3.2
direct-SCF	5.3
second order SCF	5.3.1
complete active-space SCF (CASSCF)	4.2.1
super-CI	4.2.2
full-CI	4.2.3
singly and doubly excited CI (SDCI)	4.2.4
multi-reference CI (MRCI)	4.2.5
conventional-CI	4.2.6 & 6.1.2
direct-CI	4.2.6 & 6.1.3

Table 4.1: Where chemistry methods are discussed.

In the next chapter we return to the SCF method and look at different ways of improving it.

Chapter 5

Improving the SCF Method

The SCF method has problems with convergence, speed and storage. Here we address these problems. In the previous chapter we looked at improving the accuracy and usefulness of the results by using different methods.

The energy may oscillate or increase as the iterations are performed. Thus several methods have been devised to improve the convergence of the SCF method. There is the level-shift technique and direct inversion in the iterative subspace. We look at these in sections 5.1 and 5.2 respectively. Level-shifting is used to ensure convergence and direct inversion in the iterative subspace is used to speed it up. The two techniques can be used together. The scheme described in [46] is a recent use of both these techniques.

Now some points to do with storage and parallelization are mentioned.

In general *ab initio* calculations involve far more two-electron integrals than can be stored in the central processor unit (CPU) of a computer. In the real SCF case if there are N_{BF} basis functions then there will be $\sim \frac{1}{8} N_{\text{BF}}^4$ different two-electron integrals. This is not as bad as it seems because many integrals will be effectively zero for large molecules as the distance between basis functions becomes larger. A number of integrals may also be zero because of molecular symmetry.

A given two-electron integral $(\mu\nu|\lambda\sigma)$ will be needed for six distinct elements $F_{\mu\nu}$, $F_{\lambda\nu}$, $F_{\mu\lambda}$, $F_{\mu\sigma}$, $F_{\nu\lambda}$ and $F_{\nu\sigma}$. So a program can be designed so that instead of computing each element of F one at a time, a batch of two-electron integrals is read into the CPU and then the contributions to all matrix elements that require those integrals are computed.

However transfer of the integrals between disk and CPU is relatively time consuming. This is the input/output (or I/O) of integrals. In some circumstances it is preferable to recompute the integrals in each iteration rather than transfer them from disk to CPU. This has resulted in the **direct self-consistent field** (direct-SCF) method. Direct-SCF eliminates the file space needed to store the integrals. Computations with several hundred or even a few thousand basis functions can be done. Exploiting the sparseness of the repulsion integral "matrix" is an active area of research. If sparseness and cutoff thresholds are used effectively, direct-SCF procedures are more efficient than conventional I/O based procedures, as long as the basis set is large enough [38]. This is despite the all the redundant integral computations. For large basis sets, with effective procedures, the effort for the direct-SCF calculations is $O(\sim N_{\text{BF}}^3)$. For small bases the effort is $O(N_{\text{BF}}^4)$. We use $O(N_{\text{BF}}^{<4})$ to denote this dependency.

Until recently parallel computers were not able to give chemists greater enough performance to be applicable to larger calculations, and be cheaper for computations [14]. Now work-station clusters are being used and parallel computers are more available, so more use of this technology is being made by chemists. Direct-SCF algorithms have a diagonalization bottleneck when they are parallelized. We look at a second-order scheme in section 5.3 that avoids the need for a diagonalization step. The diagonalization is replaced with a matrix inversion, matrix multiplication and additional construction of Fock-like matrices. All of these can be parallelized.

We first look at the two techniques mentioned that can be used to improve the SCF method as it is.

5.1 Level-Shifting Method

Level-shifting was due to Saunders and Hillier in 1973 [34]. It is still used in a lot of schemes. It is a way of making sure we get convergence, but it does not necessarily make it speedy.

It was first designed for intrinsically divergent wave functions. When solving the Hartree-Fock equations using the SCF method there can be problems with convergence. If a poor initial wave function is chosen we can get an oscillating or divergent sequence of energy values. A second kind of difficulty can arise when it is impossible to specify a starting set of MO's that will allow for convergence. We call this situation **intrinsically divergent**. Even if we are very close to the solution one iteration can give a wave function of higher energy than the initial one. This sort of problem is more common for the open-shell case than the closed-shell case we have looked at in detail.

Suppose the linearly independent atomic orbitals are real and orthonormal. This is without loss of generality because it is easy to get a transformation matrix to make the atomic orbitals orthonormal. Suppose we have got an N–electron system and we have $m_1 = \frac{N}{2}$ doubly occupied MO's and $m_2 = N_{\text{BF}} - m_1$ are virtual MO's. The total set of N_{BF} MO's is orthonormal.

Let $\boldsymbol{\phi}$, $\boldsymbol{\psi}$, $\boldsymbol{\psi}_1$ and $\boldsymbol{\psi}_2$ be row vectors. They contain all the AO's, all the MO's, doubly occupied MO's and virtual MO's respectively. They are related by the partitioned unitary matrix C

$$\boldsymbol{\psi} = [\boldsymbol{\psi}_1|\boldsymbol{\psi}_2] = \boldsymbol{\phi}\,(C_1|C_2) = \boldsymbol{\phi}\,C \tag{5.1}$$

where obviously

$$\boldsymbol{\psi}_1 = \boldsymbol{\phi}\,C_1 \qquad \text{and} \qquad \boldsymbol{\psi}_2 = \boldsymbol{\phi}\,C_2$$

The only type of mixing that changes the energy is between the occupied and virtual orbitals. Remembering the C is unitary here, this is consistent with what was said in subsection 2.2.4.

Let $\boldsymbol{\psi}_1^{'}$ be an improved set of occupied MO's given by

$$\boldsymbol{\psi}_1^{'} = [\boldsymbol{\psi}_1|\boldsymbol{\psi}_2] \left(\frac{I}{\Delta}\right) = [\boldsymbol{\psi}_1|\boldsymbol{\psi}_2]\,Q_1 \tag{5.2}$$

The identity I is of size m_1 and Δ is $m_2 \times m_1$. The matrix Δ is arbitrary with "small" elements. The MO's in $\boldsymbol{\psi}_1^{'}$ are not orthonormal in second and higher orders. As we only want the energy expansion to the first order in the elements of Δ we can neglect this lack of orthonormality. The elements of Δ make up the parameter space we will use and it consists of $m_1 m_2$ variables.

The change in electronic energy can be written as

$$E^{'} = E + 4 \sum_{\substack{r \\ \text{virt.}}} \sum_{\substack{a \\ \text{occ.}}} \Delta_{ra}\, F_{ra}^{[\text{MO}]} + \text{higher-order terms} \tag{5.3}$$

where $F_{ra}^{[\text{MO}]}$ is the matrix element connecting the rth virtual MO with the ath occupied MO,

$$F_{ra}^{[\text{MO}]} = \int \psi_r^*(\mathbf{r}_1) f(\mathbf{r}_1) \psi_a(\mathbf{r}_1) d\mathbf{r}_1$$

Recall that in equation (3.25) of chapter 3 we had the Fock matrix in terms of the AO's rather than the MO's as we have here. We denote the Fock matrix we had before by $F^{[\text{AO}]}$, and we have the relationship

$$\begin{aligned} F^{[\text{MO}]} &= (C_1|C_2)^T F^{[\text{AO}]} (C_1|C_2) \\ &= \left(\begin{array}{c|c} C_1^T F^{[\text{AO}]} C_1 & C_1^T F^{[\text{AO}]} C_2 \\ \hline C_2^T F^{[\text{AO}]} C_1 & C_2^T F^{[\text{AO}]} C_2 \end{array} \right) \end{aligned} \tag{5.4}$$

We need to have the diagonal blocks of $F^{[\text{MO}]}$ in diagonal form. We can assume this is true because two unitary transformations, one acting on C_1 and the other on C_2, will do the diagonalizing. Once this is true the trial occupied and virtual orbitals are *pseudo-canonical.* Looking back at equation (5.1) we see that such transformations will leave C unitary and will not change the energy, because there is no mixing between occupied and virtual orbitals. We really only need this for the derivation of the scheme. It is not part of the implementation.

There are a couple of considerations in choosing the value of Δ_{ra}. To start with we want it to have opposite sign to $F_{ra}^{[\text{MO}]}$. We also want it small enough so that the higher-order terms of equation (5.3) have an absolute value lower than that of the first term. If both of these things are true the energy of the perturbed wave function will not be any higher than that of the initial wave function.

Now consider one cycle of Roothaan's SCF procedure, which is described on page 43. We have assumed orthonormality of the basis functions or AO's so that $S=I$. This means the eigenproblem that has to be solved is

$$F^{[\text{AO}]} C' = C' \varepsilon \tag{5.5}$$

C' is the (hopefully) improved MO coefficient matrix. The m_1 eigenvectors of lowest energy define the improved occupied MO's and we can partition C' as

$$C' = \left(C_1' | C_2' \right) \tag{5.6}$$

We can diagonalize equation (5.5) in the basis of trial MO's

$$F^{[\text{MO}]} Q = Q \varepsilon \tag{5.7}$$

We can use Q to get the improved set of MO coefficients C' of equation (5.5)

$$C' = \left(C_1'|C_2'\right) = (C_1|C_2)\,(Q_1|Q_2) \tag{5.8}$$

This is consistent with the notation used in equation (5.2).

Here is where the shifting comes in. Rather than diagonalize $F^{[\mathrm{MO}]}$ we diagonalize a matrix identical to $F^{[\mathrm{MO}]}$ except that an arbitrary large positive constant γ has been added to the diagonal elements of the block $C_2^T F^{[\mathrm{AO}]} C_2$ in $F^{[\mathrm{MO}]}$. So we use the matrix

$$\left[\begin{array}{c|c} C_1^T F^{[\mathrm{AO}]} C_1 & C_1^T F^{[\mathrm{AO}]} C_2 \\ \hline C_2^T F^{[\mathrm{AO}]} C_1 & C_2^T F^{[\mathrm{AO}]} C_2 + \gamma I \end{array}\right]$$

γ is the **level-shift parameter**. Recall that the trial orbitals are pseudo-canonical, so that the diagonal blocks of the matrix we are going to diagonalize are in diagonal form.

If γ is sufficiently large Q_1 can be written in the form of equation (5.2) where

$$\Delta_{ra} = \frac{F_{ra}^{[\mathrm{MO}]}}{F_{aa}^{[\mathrm{MO}]} - F_{rr}^{[\mathrm{MO}]} - \gamma} \tag{5.9}$$

We will not derive this formula, however we will look at why it works with a small example. This gives us a way of getting values for Δ_{ra} that are arbitrarily small in magnitude. The value of $F_{rr}^{[\mathrm{MO}]}$ should be greater than all the $F_{aa}^{[\mathrm{MO}]}$ when close to the SCF solution. If not we can reorder the rows and columns in $F^{[\mathrm{MO}]}$ so that all the $F_{rr}^{[\mathrm{MO}]}$ are greater than all the $F_{aa}^{[\mathrm{MO}]}$. However after the next diagonalization this might need to be done again. This could make the procedure oscillate with the MO's being swapped from the virtual to the occupied shell. If we do this reordering and choose γ sufficiently large no extensive swapping of MO's can occur. So one iteration of a sufficiently level-shifted Roothaan process will therefore give an output wave function whose energy is not higher than that of the input wave function. This ensures that the energy decreases with each iteration and converges to a stationary point but it will not necessarily be the minimum energy.

We now look at a small example so we can see how the proof of equation (5.9) can be constructed. Suppose we have two occupied and two virtual orbitals. We want to show that Q_1 can be written in the form of equation (5.1) provided γ is sufficiently large. Assume Q_1 has this form. The eigenvalue equation for the occupied orbitals is

$$\begin{pmatrix} F_{11} & 0 & F_{31} & F_{41} \\ 0 & F_{22} & F_{32} & F_{42} \\ F_{31} & F_{32} & (F_{33}+\gamma) & 0 \\ F_{41} & F_{42} & 0 & (F_{44}+\gamma) \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ \Delta_{31} & \Delta_{32} \\ \Delta_{41} & \Delta_{42} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ \Delta_{31} & \Delta_{32} \\ \Delta_{41} & \Delta_{42} \end{pmatrix} \begin{pmatrix} \varepsilon_1 & 0 \\ 0 & \varepsilon_2 \end{pmatrix}$$

If we look at the four equations for the first eigenvalue we get

$$\begin{aligned} F_{11} - \varepsilon_1 + \frac{(F_{31})^2}{(F_{11} - F_{33} - \gamma)} + \frac{(F_{41})^2}{(F_{11} - F_{44} - \gamma)} &= 0 \\ \frac{F_{32}F_{31}}{(F_{11} - F_{33} - \gamma)} + \frac{F_{42}F_{41}}{(F_{11} - F_{44} - \gamma)} &= 0 \\ F_{31} - \frac{(\varepsilon_1 - F_{33} - \gamma)F_{31}}{(F_{11} - F_{33} - \gamma)} &= 0 \\ F_{41} - \frac{(\varepsilon_1 - F_{44} - \gamma)F_{41}}{(F_{11} - F_{44} - \gamma)} &= 0 \end{aligned}$$

By letting $\gamma \to \infty$ we see that all these equations are satisfied with $\varepsilon_1 = F_{11}$. This means if we make γ sufficiently large the equations are satisfied to any specified accuracy. It is easy to see how we can generalize this idea and apply it to bigger matrices.

Let us summarise the **level-shifting procedure**. The iterations are done in the way recommended by Roothaan, which is described in section 3.2, except the diagonalization of the Fock matrix is replaced by

(1) Compute $F^{[\text{MO}]}$ from $F^{[\text{AO}]}$ using equation (5.4).

(2) Add the pre-determined level-shift parameter γ to the diagonal elements of $C_2^T F^{[\text{AO}]} C_2$.

(3) Diagonalize the level-shifted Fock matrix and construct Q.

(4) Using equation (5.8) get the improved MO's given by C'.

To make this procedure particularly insensitive to round-off error the columns of C' can be orthonormalized using the Gram-Schmidt method for example. They should be orthonormal anyway. $F^{[\text{MO}]}$ approaches diagonal form as convergence is approached. The energies of the virtual MO's will be displaced by γ and this needs to be allowed for at the end of the iterations. Level-shifting can also be used in the open-shell situation with the RHF method, but the RHF case is more complicated.

Saunders and Hillier [34] found early on in the calculation a relatively large γ can be used, but after a few cycles this needs to be reduced otherwise convergence is slowed down.

5.2 Direct Inversion in the Iterative Subspace

This method was designed by Pulay and came to life in 1980 [28], and an improved version was presented in 1982 [27]. It is abbreviated as DIIS. It can be described as a general least squares interpolation convergence acceleration method. It can be applied to any iterative optimization scheme where an estimate of the error can be obtained at every step.

Pulay [28] pointed out that in the SCF iterative scheme the energy is essentially minimized by a quasi Newton-Raphson procedure. A consequence of this is that the convergence is approximately linear. The DIIS method is a combination of iterative and direct methods. We can use an iterative method initially and then using a direct method look for the solution in the subspace spanned by consecutive iterated vectors. With the problems we are looking at here the iterative method is the SCF procedure. DIIS achieves quadratic convergence on a quadratic surface because it exploits second-order information contained in a set of gradients. For this reason it is particularly good towards the end of the SCF procedure. It is important to have an accurately converged SCF wave function in a lot of applications. It works particularly well if the number of parameters is so large that the calculation and storage of the Hessian matrix is not practical which is the case with the SCF procedure. It is widely used and is now a standard option in most *ab initio* computer programmes.

The DIIS algorithm is a derivative of the conjugate-gradient method. Therefore it could be simplified by using a three-term reccurence relation. It was found that the use of the three-term recursion formula slowed down the convergence for this sort of problem which is highly nonlinear [13]. For this reason the three-term recursion formula does not get used with DIIS. At the end of the next chapter we will look at a case where the formula is worth using.

At the end of this section we will take a look at the C^2-DIIS algorithm which is an improved version of DIIS from 1993.

5.2.1 The DIIS Equations

We now look at the equations that we solve at each iteration of the algorithm so we can see what it is doing.

In the closed-shell SCF problem we are looking at, the product of the number of occupied and virtual orbitals gives the number of parameters. As we have already mentioned this is due to the fact that only rotations between the occupied and unoccupied orbitals change the energy. The appropriate elements of the density matrix or the Fock matrix tend to be used as the parameters.

In this chapter we will be interchanging between vectors and matrices. For example the matrix P (or p) corresponds to the vector $\mathbf{p}$ with the index association $P_{ab} = \mathbf{p}_{(ab)}$, so (ab) is treated as a single index. This means our vectors have a different numbering of elements from usual. If P was 3×3 say, then $\mathbf{p}$ would have its elements numbered 11, 12, 13, 21, 22, 23, 31, 32, 33. Chemists tend to print all their vectors, matrices and arrays in bold type. We do not use their convention so that it is clearer when we have swapped between a vector and a matrix, and vice versa.

The energy $E = E(p_1, p_2, \ldots, p_n)$ can be considered to be quadratic in the parameters p_i that are being varied when it is close to convergence. We assume that we have a trial set of parameters $\mathbf{p}^{(1)} = (p_1^{(1)}, p_2^{(1)}, \ldots, p_n^{(1)})^T$ already and that we are close enough to the final solution $\mathbf{p}^{\mathrm{f}}$. Then according to the quasi Newton-Raphson method an improved set $\mathbf{p}^{(2)}$ will be given by

$$\mathbf{p}^{(2)} = \mathbf{p}^{(1)} - H_0^{-1}\mathbf{g}^{(1)}$$

where $\mathbf{g}^{(1)}$ is the gradient vector at $\mathbf{p}^{(1)}$, and H_0^{-1} is an approximate inverse Hessian. Let the exact Hessian be given by H. Within the quadratic region $\mathbf{g} = H(\mathbf{p} - \mathbf{p}^{\mathrm{f}})$ so that

$$\mathbf{p}^{(2)} = \mathbf{p}^{(1)} - H_0^{-1}H\left(\mathbf{p}^{(1)} - \mathbf{p}^{\mathrm{f}}\right)$$

$$= \mathbf{p}^{\mathrm{f}} + \left(I - H_0^{-1}H\right)\left(\mathbf{p}^{(1)} - \mathbf{p}^{\mathrm{f}}\right)$$

We can replace $\mathbf{p}^{(1)}$ with $\mathbf{p}^{(2)}$ and this can be iterated. We will call this iterative technique **simple relaxation** (SR). If $H_0 = H$ then SR will converge in one step within the quadratic region.

As it is not practical to evaluate and invert H we do something else. In the SCF case an acceptable approximation is available. If the eigenvalues of $I - H_0^{-1}H$ are less than 1 in magnitude SR converges. However it can be really slow unless H_0 is a good approximation to H, with the eigenvalues of $I - H_0^{-1}H$ being much less than 1 in magnitude. We get a much better approximation to $\mathbf{p}^{\mathrm{f}}$ using the following procedure.

The method we are going to look at constructs a new $\mathbf{p}$ in terms of the $\mathbf{p}^{(i)}$'s from iterations 1 to m

$$\mathbf{p} = \sum_{i=1}^{m} a_i \mathbf{p}^{(i)} \tag{5.10}$$

Let $\mathbf{e}^{(i)}$ be the error vector for the ith iteration. The a_i are chosen so that the **error vector**

$$\mathbf{e} = \sum_{i=1}^{m} a_i \mathbf{e}^{(i)} \tag{5.11}$$

approaches the zero vector using the two-norm and such that

$$\sum_{i=1}^{m} a_i = 1 \tag{5.12}$$

This means that the trivial solution with all the coefficients equal to zero will not be possible. We can use different definitions of the error vector. This means we can look for the convergence of different quantities.

Towards the end of the SCF procedure changes in the wave function are small and we can assume the error vector depends linearly on the parameters. In fact in most applications the error quantities are some sort of gradient, and the gradients vary linearly with the independent variables near the solution. This is the reason for minimizing a function given by (5.11).

The Lagrangian for this problem is

$$\begin{aligned}\mathcal{L}(\mathbf{a}, \lambda) &= \|\mathbf{e}\|_2^2 - 2\lambda\left(\sum_{i=1}^{m} a_i - 1\right)\\ &= \langle \mathbf{e}|\mathbf{e}\rangle - 2\lambda\left(\sum_{i=1}^{m} a_i - 1\right)\end{aligned}$$

$$= \sum_{j=1}^{m}\sum_{i=1}^{m} a_i a_j \langle \mathbf{e}^{(i)}|\mathbf{e}^{(j)}\rangle - 2\lambda\left(\sum_{i=1}^{m} a_i - 1\right) \tag{5.13}$$

We minimize $||\mathbf{e}||_2^2$ rather than $||\mathbf{e}||$ because it gives a simpler equation. Also we have 2λ rather than λ so that we have λ rather than $\frac{\lambda}{2}$ in the matrix equation (5.15) below.

The conditions $\frac{\partial \mathcal{L}}{\partial a_j} = 0 \;\; j = 1, \ldots, m$ give

$$\sum_{i=1}^{m} a_i \langle \mathbf{e}^{(i)}|\mathbf{e}^{(j)}\rangle - \lambda = 0 \qquad\qquad j = 1, \ldots, m \tag{5.14}$$

This together with $\frac{\partial \mathcal{L}}{\partial \lambda} = 0$ means that the a_i are calculated from

$$\begin{bmatrix} B_{11} & B_{12} & \cdots & B_{1m} & -1 \\ B_{21} & B_{22} & \cdots & B_{2m} & -1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ B_{m1} & B_{m2} & \cdots & B_{mm} & -1 \\ -1 & -1 & \cdots & -1 & 0 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \\ \lambda \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ -1 \end{bmatrix} \tag{5.15}$$

where

$$B_{ij} = \mathbf{e}^{(i)T}\mathbf{e}^{(j)} = \langle \mathbf{e}^{(i)}|\mathbf{e}^{(j)}\rangle \tag{5.16}$$

and λ is a Lagrange multiplier. From equations (5.14) we have m different ways of expressing λ. Using these and the condition $\sum_{j=1}^{m} a_j = 1$ we get

$$\lambda = \sum_{j=1}^{m} a_j \lambda = \sum_{j=1}^{m} a_j \left(\sum_{i=1}^{m} a_i \langle \mathbf{e}^{(i)}|\mathbf{e}^{(j)}\rangle\right) = \langle \mathbf{e}|\mathbf{e}\rangle$$

So the Lagrange multiplier is equal to the quantity we are minimizing. Therefore the Lagrange multiplier gives us an indication of how good a solution is.

We call equations (5.15) the **DIIS equations**. They are the basis for the DIIS method which we will soon outline.

5.2.2 The Error Vector (Matrix)

In [28] the error vector, which is the quantity we are trying to get to converge to zero, was the change in the parameter vector in the course of the subsequent SCF step, ie.

$$\mathbf{e}^{(i)} = \Delta\mathbf{p}^{(i)} = \mathbf{p}^{(i+1)} - \mathbf{p}^{(i)}$$

This is what we might expect the error to be. This choice worked well in the semi-empirical applications that DIIS was originally developed for. However it was found to have a few weaknesses.

In the Roothaan SCF procedure the main computational task is the construction of the Fock matrix. In the first algorithm [28] the determination of the error vector needed the construction of an extra Fock matrix and this Fock matrix was not used in any other way. Because of the large number of arithmetic operations needed to construct the error vector it was not sufficiently accurate numerically. The main problem was that the DIIS step was performed only periodically after an arbitrary (usually 5–12) number of conventional SCF cycles. This restricted the use of DIIS in intrinsically divergent or very slowly convergent cases.

In the new method [27] the DIIS equations are solved at every step. The new method is also based on a new definition or choice of the error vector. This avoids the previous deficiencies.

A necessary and sufficient condition for SCF convergence is the vanishing of the off-diagonal blocks of the Fock matrix in the MO basis. When this happens there is no mixing between different orbital groups. This is what the new definition of the error vector is based on. The orbitals are grouped together according to occupation number. So in the closed-shell case there are two different orbital groups, and in a simple restricted open-shell wave function there are three groups. The density matrix below is more general than the charge density matrix $D = 2CC^T$ we have been considering, but it includes this simple case. We end up with an error matrix [27]

$$e \equiv FDS - SDF \tag{5.17}$$

The density matrix D in this formula is the one used to construct the Fock matrix F that appears in the formula.

It was pointed out back in 1969 that we can get a SCF solution if and only if the density matrix commutes with the Fock matrix, and this is also true in the open-shell case. The error definition makes sense in terms of this.

A more *balanced* error vector can be obtained by transforming e to an orthonormal basis. This is equivalent to replacing e by $e' = A^TeA$ where $A^TSA = I$, for example $A = S^{-\frac{1}{2}}$. This is used in the algorithm in the next subsection.

5.2.3 The DIIS Algorithm

The **direct inversion in the iterative subspace procedure** (with the Fock matrix elements as parameters) is:

(1) Obtain a starting vector $\mathbf{p}^{(1)}$ which is given by a Fock matrix F.

(2) Construct the error vector $\mathbf{e}$ according to (5.17) and transform to an orthonormal basis. If the largest element of the error matrix $e_{\max}$ is less than a threshold ($\simeq 0.1\,E$, where E is approximately equal to the converged energy) initiate DIIS. Otherwise use a SCF iteration. If $e_{\max}$ is less than another threshold ($\simeq 10^{-6}\,E$) the SCF procedure has converged.

(3) From now in each DIIS step store the error matrix and the current Fock matrix. Evaluate the necessary scalar products $B_{ij} = \mathrm{Tr}\,(e_i\, e_j^T)$. (We have used the index association $\mathbf{e}_{(ab)} = e_{ab}$ and this is just $\mathbf{e}_i^T\mathbf{e}_j$.)

(4) Set up and solve equation (5.15). If equation (5.15) becomes ill-conditioned omit the first, second, ... row and column until its condition becomes acceptable. This corresponds to truncating the DIIS space.

(5) Replace the current Fock matrix by $F' = \sum_{i=1}^{m} a_i F_i$.

(6) Do a SCF-type iteration; diagonalize F', determine the new density matrix and the new Fock matrix. Go back to step (2).

Note that in step (3) it is only necessary to evaluate the last row of the matrix B in each iteration.

Also note that DIIS violates the idempotency of the density matrix but this is insignificant as convergence is approached. In the closed-shell case a sufficiently good starting wave function is usually available, and in an orthonormal basis it is $\frac{1}{2}D$ that is idempotent.

DIIS can converge in principle even if the underlying SR procedure diverges, but it is difficult to get to and remain in the quadratic region when this is the case. This is the case with intrinsically divergent wave functions. The convergence properties of DIIS are discussed in [28] and compared with that of SR. According to Pulay [27] this method is greatly superior to the SCF procedure particularly in the final stages when convergence is usually slow. There are several reasons why DIIS is suited to accelerating the convergence of the SCF procedure. Firstly the parameters do not need to be independent for the algorithm to be used and redundant parameters should not change the convergence. However dependencies can cause other numerical problems. The gradient vector does not need to be evaluated explicitly which does require independent parameters. The gradient and approximate Hessian are implicit in the SCF procedure and that is all that is needed. Usually the maximum element of the error vector decreases almost an order of magnitude per step towards the end of the iteration if DIIS is used, compared with a decrease by a factor of about 2 in ordinary SCF. The extra computational effort and storage requirements associated with DIIS are negligible.

In several SCF applications a higher degree of convergence is needed and this is what DIIS was designed for. However it is of little or no use in getting initial convergence. One of the applications is orbital generation for large-scale CI calculations.

The details of the implementation of DIIS depends on the actual problem that is being solved. Császár and Pulay [4] use geometry displacements as the error quantities in their GDIIS algorithm. They are trying to locate a stationary point on a nearly quadratic potential energy surface. Sellers [37] uses DIIS with an iterative subspace that is generated by a direct energy minimization procedure. It can be thought of as an improvement on DIIS because it forces the iterative subspace to contain more of the final solution. A very recent use of DIIS was in the method proposed in [46]. It used Fock matrices in the DIIS method and hence formed a linear combination of Fock matrices to construct an improved SCF wave function.

DIIS can be extended to more general wave functions. This is done by Hamilton and Pulay in [13].

5.2.4 The C^2-DIIS Algorithm

The C^2-DIIS algorithm is intended to allow the use of larger DIIS spaces without giving up numerical stability. It was published by Sellers in 1993.

The DIIS algorithm has numerical problems when near linear dependencies build up in the error or gradient vector space. This often happens when the dimension of the DIIS space is allowed to become large, and too often even in small dimension problems. The errors show up as large coefficients a_i which can result when the matrix in equation (5.15) is nearly singular. The nonlinearities of the scalar field E can be magnified by the large coefficients. This ruins the relationship between equations (5.10) and (5.11), because quadratic behaviour was assumed when the equations were derived. Whether this happens depends on the nature of the energy field. Round-off errors can accumulate and swamp the true solution. This probably happens more often in practice when the dimension of the DIIS space is allowed to get large.

To get the DIIS equation given by (5.15) we required that the sum of the coefficients a_i be one. For this reason Sellers [36] uses the notation C^1-DIIS for the DIIS algorithm given in subsection 5.2.3.

The C^2-DIIS algorithm that is about to be described is equivalent to the C^1-DIIS algorithm near convergence. It is more stable numerically so that round-off errors do not accumulate as much. Consequently it can be used with much larger DIIS spaces and the DIIS spaces do not need to be truncated.

Sellers C^2-DIIS algorithm [36] is obtained in much the same way as the C^1-DIIS algorithm. The constraint that the DIIS coefficients add to one is replaced with the constraint that the squares of the coefficients sum to unity. This is the reason for the C^2-DIIS notation. A renormalization is done at the end so that the coefficients sum to one. The Lagrangian is now

$$\mathcal{L}(\mathbf{a}, \lambda) = \sum_{j=1}^{m} \sum_{i=1}^{m} a_i a_j \langle \mathrm{e}^{(i)} | \mathrm{e}^{(j)} \rangle - \lambda \left(\sum_{i=1}^{m} a_i^2 - 1 \right) \tag{5.18}$$

The conditions $\frac{\partial \mathcal{L}}{\partial a_j} = 0$ give

$$B\mathbf{q}_j = \lambda_j \mathbf{q}_j \tag{5.19}$$

where $B_{ij} = \langle \mathrm{e}^{(i)} | \mathrm{e}^{(j)} \rangle$ and $\mathbf{q}_j$ is the jth eigenvector of B with the a_i as its elements. The eigenvalue λ_j is the Lagrange multiplier which is indexed with j because it can take m different values. The condition $\frac{\partial \mathcal{L}}{\partial \lambda} = 0$ gives the normalization condition. To get properly normalized DIIS coefficients $\mathbf{q}_j$ is divided by the sum of its elements which we denote by n_j.

Obviously there are many possible solution vectors. This multiplicity of C^2-DIIS is an advantage when linear dependencies start to build up. When elements of $n_j^{-1}\mathbf{q}_j$ become big enough to exaggerate the nonlinearities in E there will be other solutions to choose from. To choose a solution from all the possible $n_j^{-1}\,\mathbf{q}_j$ we evaluate equation (5.11) for each of them. We then choose the one that gives the error of the smallest size. It must also have elements with magnitudes below some threshold that has already been specified, and it must still be large enough that it is not contaminated by round-off error. We do not throw out a C^2-DIIS solution if its corresponding eigenvalue is zero or about the size of the expected round-off. The best solution gives a zero value of $\lambda = \langle \mathbf{e}|\mathbf{e}\rangle$. It can be shown that this quantity is related to the lowest eigenvalue of B, and that the C^2-DIIS algorithm gives the same solution as the original C^1-DIIS method when the lowest eigenvalue of B approaches zero [36, pages 34–35]. This is when we are near to convergence.

A null value of $\langle \mathbf{e}|\mathbf{e}\rangle$ can be thought of as an allowed or desired linear dependence. It seems that the C^2-DIIS algorithm might not be able to distinguish between the allowed solution and ordinary linear dependences that must eventually build up in the DIIS space. This is not a problem in practice [36], because of the criterion that is used for throwing out solutions.

Reference [36] gives an example that illustrates the numerical similarities between the two procedures. In Sellers' [36] experience C^2-DIIS performs as well as C^1-DIIS in well-behaved cases and offers alternative solutions in situations in which the usual DIIS equations are ill-conditioned. The C^2-DIIS method can be thought of as a variant of the singular value decomposition method because both eliminate linear dependencies by using numerically stable matrix diagonalization techniques.

5.3 Parallel Direct Second-Order SCF Methods

We now look at the methods proposed by Shepard in 1993 [38]. The second-order iterative procedure that Shepard's idea is based on has better convergence properties and avoids the need for a diagonalization step.

First we mention the work needed. In sequential direct-SCF procedures the diagonalization of the Fock matrix at each iteration is a bottleneck. This takes

$O(N_{\mathrm{BF}}^3)$ work. In Shepard's method this is replaced with a combination of parallel $O(N_{\mathrm{BF}}^{\leq 4})$ and sequential $O(N_{\mathrm{Sub}}^3)$ steps. N_{Sub} is the dimension of the expansion subspace, and $N_{\mathrm{Sub}} \ll N_{\mathrm{BF}}$ for large basis sets.

The equation $F[C]C = SC\varepsilon$ comes from the three independent conditions:

- The molecular orbitals remain orthonormal.
- The total energy is stationary with respect to orbital variations.
- The orbitals are chosen to be in their canonical form.

The first two conditions are enforced in subsection 2.2.1. If we think of it as an optimization problem, the first condition is the constraint functions and the second condition is the function that is minimized. The third comes in when the Hartree-Fock equations are put into canonical form in subsection 2.2.4. So traditional SCF methods simultaneously optimize and canonicalize the occupied orbitals.

In more recent MCSCF theories, where the role of the Fock operator is less important, these three conditions have been obtained separately from each other. Orthonormalization is obtained by choosing the right wave function variation coordinates, and canonicalization (of some sort) is only applied as an after thought if at all. While the orbital optimization equation is solved orbital canonicalization does not come into it. Instead the expectation value of the energy is regarded as a function of all possible orbital variations, and the energy is minimized within this variational space. This view point can be applied to single configuration wave functions too. We will be doing this here so canonicalization will not come into the method. It can be done as an extra calculation after convergence has been reached.

There are two different things that can be considered more basic to the SCF wave function. The first is the minimization of the energy with respect to orbital variations. The other is determining a self-consistent Fock operator.

Now we turn to the energy point of view to show that we can get an iterative method that does not involve matrix diagonalization. The reason for doing this is diagonalization is particularly difficult to programme efficiently in parallel. This is true on either distributed or shared memory machines, and on either single-instruction-multiple-data (SIMD) or multiple-instruction-multiple-data (MIMD) architectures. This is due to the compromises that have to be made between load balancing and the communication overhead. Parallel matrix computations are discussed in [11, pages 275–307]. For discussions more specific to what we are looking at here, see Mattson [23], Harrison and Shepard [14], and Kendall, Harrison, Littlefield and Gress in chapter four of [22].

The idea behind what we are going to look at applies to closed, open, restricted and unrestricted wave functions. For simplicity we continue to focus on closed-shell restricted wave functions.

5.3.1 Second Order SCF Methods

Bacskay has developed a second-order SCF method for sequential implementation [1]. We look at this now for a closed-shell restricted Hartree-Fock wave function. What we do is optimize without canonicalization.

For now we assume an orthonormal MO basis $\{\psi_i\}$ is available, and later we write the computational steps in the AO basis $\{\phi_\mu\}$. Put the bases into row vectors and assume they are related by the matrix C,

$$\psi = \phi C$$

A transformation to a new orthonormal MO basis ψ' can be written as

$$\psi' = \psi U$$

where U is a real rotation matrix, so that $U^T U = I$ and $\det(U) = 1$. We can parameterize U as $U = \exp(K)$ with $K = -K^T$. (This matrix K is different from the K used for exchange integrals in equation (3.11).) The occupied orbitals are grouped together so that the matrix K is partitioned into four blocks like this

$$K = \left[\begin{array}{cc|cc} K_{ab} & \cdots & K_{ar} & \cdots \\ \vdots & \ddots & \vdots & \ddots \\ \hline -K_{ra} & \cdots & K_{rs} & \cdots \\ \vdots & \ddots & \vdots & \ddots \end{array}\right]$$

Let us look at why this parameterization can be given this physical significance. It is obvious that this sort of physical significance could be given to the elements of U but not so obvious for that of K. Since U is a real rotation matrix it can be expressed as $U = VSV^T$. The matrix V is orthogonal and S is block diagonal with 1's and 2×2 blocks of the form

$$\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

on the diagonal. We have $S = \exp\Lambda$ where Λ is block diagonal with 0's (corresponding to the 1's in S) and 2×2 blocks of the form

$$\begin{bmatrix} 0 & -\theta \\ \theta & 0 \end{bmatrix}$$

(corresponding to the 2×2 blocks in S) on the diagonal. The matrix V changes to an orthonormal basis where U has the more simple form S. Now

$$U = V \exp\Lambda\, V^T = \exp(V\Lambda V^T)$$

It is the matrix $V\Lambda V^T$ that corresponds to K.

The essential, or non redundant, parameters occur only in the off-diagonal block of K. These elements are K_{ar}, where a is the occupied subscript and r is the unoccupied subscript. The other elements of the matrix, the K_{ab} and K_{rs}, which appear in the diagonal blocks, do not change the energy. Note that they are grouped together in K not U. So we can set the K_{ab} and K_{rs} to zero without losing any generality. This corresponds to setting the appropriate angle θ to zero in the matrix Λ.

It is an appropriate time to mention the indices that will be used. We use a and b for occupied orbitals, and r and s for unoccupied orbitals. The Greek letters μ and ν will be matrix indices and will mostly be used for MO expressions, and i and j will be summation indices and will generally be for AO expressions.

The wave function variations are parameterized with the $\{K_{ar}\}$. We can expand the energy about the current wave function as

$$E(\mathbf{k}) = E(\mathbf{0}) + \mathbf{w}^T\mathbf{k} + \tfrac{1}{2}\mathbf{k}^T B\mathbf{k} + 0(K^3) + \ldots \tag{5.20}$$

The vector $\mathbf{k}$ is defined in terms of the matrix elements as $k_{(ar)} = K_{ar}$. The length of $\mathbf{k}$ is equal to the product of the number of occupied and unoccupied orbitals, ie. the number of nonzero elements of K. If $\mathbf{k} = \mathbf{0}$ then K is the zero matrix so U is the identity and we have the current wave function. The gradient vector is $\mathbf{w}$ and the Hessian matrix is B. Their elements are given by the following two equations,

$$w_{(ar)} = 2F_{ar}^{[\text{MO}]} \tag{5.21}$$

$$B_{(ar)(bs)} = 2F_{rs}^{[\text{MO}]}\delta_{ab} - 2F_{ab}^{[\text{MO}]}\delta_{rs} + 8\left(2g_{rasb}^{[\text{MO}]} - \tfrac{1}{2}g_{rbsa}^{[\text{MO}]} - \tfrac{1}{2}g_{rsab}^{[\text{MO}]}\right) \tag{5.22}$$

The Fock matrix $F^{[\text{MO}]}$, in the MO basis, is defined as

$$\begin{aligned} F_{\mu\nu}^{[\text{MO}]} &= 2h_{\mu\nu}^{[\text{MO}]} + \sum_a 4g_{\mu\nu aa}^{[\text{MO}]} - 2g_{\mu a\nu a}^{[\text{MO}]} \\ &= 2h_{\mu\nu}^{[\text{MO}]} + \sum_{ab}\left(2g_{\mu\nu ab}^{[\text{MO}]} - \tfrac{1}{2}g_{\mu a\nu b}^{[\text{MO}]} - \tfrac{1}{2}g_{\mu b\nu a}^{[\text{MO}]}\right)2\delta_{ab} \\ &= 2h_{\mu\nu}^{[\text{MO}]} + \sum_{ab}\left(2g_{\mu\nu ab}^{[\text{MO}]} - \tfrac{1}{2}g_{\mu a\nu b}^{[\text{MO}]} - \tfrac{1}{2}g_{\mu b\nu a}^{[\text{MO}]}\right)2D_{ab}^{[\text{MO}]} \end{aligned} \tag{5.23}$$

This can be compared with expression (3.25) in subsection 3.2.3. The summations are only over the occupied orbitals like before. The other notation has changed a bit. The arrays $h^{[\text{MO}]}$ and $g^{[\text{MO}]}$ are the one-electron Hamiltonian and two-electron repulsion arrays respectively. So that

$$h_{\mu\nu}^{[\text{MO}]} = \int \psi_\mu(\mathbf{r}_1)h(\mathbf{r}_1)\psi_\nu(\mathbf{r}_1)d\mathbf{r}_1$$

and

$$g_{\mu\nu\sigma\lambda}^{[\text{MO}]} = (\mu\nu|\sigma\lambda) = \int \psi_\mu(\mathbf{r}_1)\psi_\nu(\mathbf{r}_1)r_{12}^{-1}\psi_\sigma(\mathbf{r}_2)\psi_\lambda(\mathbf{r}_2)d\mathbf{r}_1 d\mathbf{r}_2$$

The operator h is defined by equation (2.17). The definition of the Fock matrix differs from that of equation (3.25) by a factor of two. The old definition is the one that gets used in SCF theory, and the new one is consistent with other areas of electronic structure theory. The matrix $D^{[\text{MO}]} = 2I$ is the density matrix in the MO basis.

We can use equation (5.23) to convert back into the AO basis.

$$F_{\mu\nu}^{[\text{MO}]} = \sum_{ij} C_{i\mu}C_{j\nu}\left(2h_{ij}^{[\text{AO}]} + \sum_{kl}\left(2g_{klij}^{[\text{AO}]} - \tfrac{1}{2}g_{kilj}^{[\text{AO}]} - \tfrac{1}{2}g_{kjli}^{[\text{AO}]}\right)D_{kl}^{[\text{AO}]}\right)$$

Here we have used the identity $g_{ijkl} = g_{klij}$, which corresponds to swapping the integration variables. The $h_{ij}^{[\text{AO}]}$ and $g_{klij}^{[\text{AO}]}$ are now in the AO basis, so $h_{ij}^{[\text{AO}]} = \int \phi_i(\mathbf{r}_1)h(\mathbf{r}_1)\phi_j(\mathbf{r}_1)d\mathbf{r}_1$. The matrix $D^{[\text{AO}]}$ is defined by

$$D^{[\text{AO}]} = 2\,C\,C^T = C\,D^{[\text{MO}]}\,C^T \tag{5.24}$$

This is consistent with the previous definition (3.23).

We can put

$$
\begin{aligned}
F^{[\mathrm{MO}]}_{\mu\nu} &= \sum_{ij} C_{i\mu} C_{j\nu} F^{[\mathrm{AO}]}_{ij} = \left(C^T F^{[\mathrm{AO}]} C\right)_{\mu\nu} \\
&= \left(C^T \left(2h^{[\mathrm{AO}]} + Q\left[D^{[\mathrm{AO}]}\right]\right) C\right)_{\mu\nu}
\end{aligned} \tag{5.25}
$$

where we have

$$
F^{[\mathrm{AO}]} = 2h^{[\mathrm{AO}]} + Q\left[D^{[\mathrm{AO}]}\right] \tag{5.26}
$$

The matrix $\frac{1}{2}F^{[\mathrm{AO}]}$ corresponds to the Fock matrix $F[C]$ used in traditional SCF optimization methods. In the last expression for $F^{[\mathrm{MO}]}_{\mu\nu}$ the one-electron and two-electron contributions have been separated. This is similar to what was done back in equation (3.25). The matrix $Q[D^{[\mathrm{AO}]}]$ is defined by

$$
Q_{ij}[D^{[\mathrm{AO}]}] = \sum_{kl} \left(2g^{[\mathrm{AO}]}_{klij} - \tfrac{1}{2}g^{[\mathrm{AO}]}_{kilj} - \tfrac{1}{2}g^{[\mathrm{AO}]}_{kjli}\right) D^{[\mathrm{AO}]}_{kl} \tag{5.27}
$$

If we optimize the $\{K_{ar}\}$ by truncating equation (5.20) at the second order term we get the Newton-Raphson procedure. This method does not get used much in MCSCF optimization because it has poor global behaviour. So alternative methods have been developed. The SCF case is discussed by Bacskay [1], and the MCSCF case is discussed in [19, pages 63–200] and [39]. In the general MCSCF case it is sufficient to work out the correction vector in terms of subspace representations of $\mathbf{w}$ and B. We will assume this is true for the simpler SCF wave functions.

A set of linearly independent vectors $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots$ are collected and form the columns of the matrix X. The subspace representations of the gradient and Hessian are given by

$$
\widehat{\mathbf{w}} = X^T \mathbf{w} \tag{5.28}
$$

and

$$
\widehat{B} = X^T B\, X \tag{5.29}
$$

respectively. The correction vector will end up being a linear combination of these trial vectors

$$
\mathbf{k} = X\, \widehat{\mathbf{k}} \tag{5.30}
$$

The subspace vector $\widehat{\mathbf{k}}$ depends on the details of the iterative procedure. For the moment we can assume that $\widehat{\mathbf{k}}$ is the solution of the subspace representation of the rational function approximation to (5.20). The subspace vector $\widehat{\mathbf{k}}$ is the solution to the eigenproblem

$$\begin{pmatrix} \widehat{B} & \widehat{\mathbf{w}} \\ \widehat{\mathbf{w}}^T & 0 \end{pmatrix} \begin{pmatrix} \widehat{\mathbf{k}} \\ 1 \end{pmatrix} = \lambda \begin{pmatrix} \widehat{\mathbf{k}} \\ 1 \end{pmatrix} \tag{5.31}$$

This choice is good enough in the MCSCF case. This is particularly true when it is combined with trust-radius and level-shifting procedures [19, pages 189–191]. However in the simpler SCF case it might not be optimal. We have already looked at level-shifting and will not be look at trust-radius methods in any detail. A trust-radius is like a trust-region in optimization. It is a measure of the size of the neighbourhood around some reference wave function within which the exact energy is adequately represented by the approximate energy expression.

5.3.2 The Parallel Method

We now look at the actual equations that are used to compute the subspace representations. After this we outline the procedure.

The matrix-vector products $\boldsymbol{\sigma} = B\,\mathbf{x}$ are needed to compute $\widehat{B}$. Let us look at this. In the MO basis we have got

$$\sigma_{(ar)} = \sum_{(bs)} B_{(ar)(bs)} x_{(bs)}$$

We treat $\boldsymbol{\sigma}$ and $\mathbf{x}$ as matrices by using $\sigma_{(ar)} \to \sigma_{ar}$, so they are the same size as K and F. We have

$$\sigma_{ar} = 2\sum_s F^{[\mathrm{MO}]}_{rs} x_{as} - 2\sum_b F^{[\mathrm{MO}]}_{ab} x_{br} + 4\sum_{bs}\left(2g^{[\mathrm{MO}]}_{rasb} - \tfrac{1}{2}g^{[\mathrm{MO}]}_{rbsa} - \tfrac{1}{2}g^{[\mathrm{MO}]}_{rsab}\right)(2x_{bs})$$

The first two terms are just simple matrix products. Efficient procedures for this are available on essentially all parallel machines. If we express the rest of the contribution in terms of the AO basis we get

$$\sigma = 2(xF^{[\mathrm{MO}]}) - 2(F^{[\mathrm{MO}]}x) + 4C^TQ[Z^{[\mathrm{AO}]}]C \tag{5.32}$$

We have used the fact the F is symmetric. The σ and x are no longer bold because they are now matrices. The intermediate matrix $Z^{[\mathrm{AO}]}$ is symmetric and defined by

$$Z^{[\mathrm{AO}]} = Y + Y^T \tag{5.33}$$

where

$$Y_{\mu\nu} = \sum_{bs} C_{\mu b} x_{bs} C_{\nu s} = \left(CxC^T\right)_{\mu\nu} \tag{5.34}$$

We can see from equation (5.32) that the two-electron repulsion integral contribution can be computed in exactly the same way as for $F^{[\text{AO}]}$ in equation (5.25). The only difference is that we have $Z^{[\text{AO}]}$ instead of $D^{[\text{AO}]}$.

After the orbital correction matrix K has been determined with an iterative procedure the transformation matrix $U = \exp(K)$ has to be computed. In the usual MCSCF procedures this is done by factoring K as $V \Lambda V^T$, where V is orthogonal and Λ is block diagonal with at most 2×2 sub-blocks. Then $U = V \exp(\Lambda) V^T$. However in parallel implementations this cannot be done effectively because it is about as hard as diagonalization. We do not have a practical alternative way of computing U exactly that avoids a diagonalization step [38].

Instead of computing U exactly we have two reasonable options. We can use the truncated expansion

$$U = I + K + \tfrac{1}{2} K^2$$

and then orthonormalize the matrix. We do not need to expand past the second order term because this is enough to guarantee overall second-order convergence in the neighbourhood of the final solution. The orthonormalization is about as expensive as diagonalization $O(N_{\text{BF}}^3)$, but it does not have the same communication and load-balancing bottleneck. Secondly we can use a rational approximation, such as

$$U = (I - \tfrac{1}{2} K)^{-1} (I + \tfrac{1}{2} K) \tag{5.35}$$

This is accurate to the second order, and it is already orthogonal. So no additional orthonormalization step is needed. It is expected that this would be the best overall approach because most interesting architectures are designed to solve linear equations at their fastest rates.

The whole **direct-SCF procedure for parallel implementation** is:

(1) Choose a starting C.

(2) Compute $D_{kl}^{[\text{AO}]} = 2 \sum_a C_{ka} C_{la}$.

(3) Compute $F^{[\mathrm{AO}]}$ using equations (5.26) and (5.27).

(4) Compute $F^{[\mathrm{MO}]}$ using equation (5.25).

(5) Put $n=1$. Choose a starting $\mathbf{x}^{(1)}$ (which is the column of X), and $n_{\max}$ which is the maximum number of iterations.

(6) Compute $Z^{[\mathrm{AO}]}$ using equation (5.34) for Y and equation (5.33) for $Z^{[\mathrm{AO}]}$.

(7) Compute $Q[Z^{[\mathrm{AO}]}]$ of equation (5.32) using equation (5.27) with $Z^{[\mathrm{AO}]}$ in place of $D^{[\mathrm{AO}]}$.

(8) Compute $\boldsymbol{\sigma}^n$ using equation (5.32).

(9) Compute $\widehat{\mathbf{w}}$, $\widehat{B}$, $\widehat{\mathbf{k}}$ and form $\mathbf{k}$.

(10) Check convergence and if converged goto step (12).

(11) Put $n=n+1$. Construct $\mathbf{x}^{(n)}$ (which is added to X to give it another column), and go back to step (6).

(12) Construct K from the final $\mathbf{k}$ $(=X\widehat{\mathbf{k}})$.

(13) Construct U from equation (5.35).

(14) Update $C \leftarrow CU$.

(15) Check convergence and if converged exit. Otherwise go back to step (2).

Steps (5) to (11) represent the subspace iterations.

Now consider how much work each of the steps take. We will ignore the work required to check convergence and in getting the initial C. The work involved in getting $\mathbf{x}^{(1)}$ and constructing $\mathbf{x}^{(n+1)}$ will also be ignored. This depends on the subspace approximation that is being used. This means steps (1), (5), (10), (11) and (15) are ignored. Step (2) requires minimal work. Step (12) does not require any additional calculations because we already have $\mathbf{k}$ from step (9). Table 5.1 shows the work required at all the other steps of the algorithm. The subspace iterations are in the middle of the table. In step (9) the work is in finding the subspace solution for

k. The eigenvector solution of equation (5.31) requires N^3_{Sub} effort, and this should be about the same as the other subspace methods. We see that the whole procedure requires parallel $O(N^{<4}_{\text{BF}})$ and sequential $O(N^3_{\text{Sub}})$ work.

step	*sequential or parallel*	*work*
3	p	$N^{<4}_{\text{BF}}$
4	p	N^3_{BF}
6	p	N^3_{BF}
7	p	$N^{<4}_{\text{BF}}$
8	p	N^3_{BF}
9	s	N^3_{Sub}
13	p	N^3_{BF}
14	p	N^3_{BF}

Table 5.1: The work taken at the relevant steps of the parallel direct-SCF algorithm.

In the usual MCSCF procedure the subspace is initialized with one vector. This is usually the gradient vector of that iteration, so $x^{(1)}_{(ia)} = w_{(ia)} = 2F^{[\text{MO}]}_{ia}$. Then each subspace iteration produces one more vector and the dimension of the subspace goes up by one at each iteration. In the SCF case there are other possibilities [38, page 349]. This is because computing several Q matrices requires insignificant extra work compared to computing one Q. We could use the correcting vector from previous SCF iterations. It is also possible to use data compression techniques with the subspace expansion vectors. Data compression for first-order iterative processes is looked at in [40].

When we use the above method we do not generally get a diagonal Fock matrix $F^{[\text{MO}]}$. However after the optimization procedure has been completed we can apply canonicalization. If it is restricted to rotations among only the occupied orbitals, or only the unoccupied orbitals, or both, it will not affect the wave function. It only changes the representation of the wave function and the energy will be unchanged. This canonicalization is not part of the iterative process. If $F[C^{(i)}]$ is the final Fock matrix, then the matrix C^{i+1}, from $F[C^{(i)}]C^{(i+1)} = SC^{(i+1)}\varepsilon^{(i+1)}$, will give a Fock matrix $F[C^{(i+1)}]$ that satisfies $FC = SC\varepsilon$ to numerical accuracy. Nothing more is needed to get self-consistency.

Chapter 6

Davidson's Method

In section 4.2 we had a brief look at configuration interaction. At the end of that section the CI method was briefly discussed. Davidson's method is used for solving the eigenvalue problem which crops up in the CI method.

We now look at CI and the CI method again. After that we look at the Lanczos algorithm and then at Davidson's method, which can be described as a combination of preconditioning and Gram-Schmidt orthogonalization with the Lanczos method. We also look at a generalized version of Davidson's method which includes the Lanczos algorithm. In the final section we look at some recent modifications of Davidson's method for CI problems. The main one that is discussed is due to van Lenthe and Pulay and can be described as a simplification of the Lanczos algorithm.

6.1 The Configuration Interaction Method

CI is used for general studies of molecular systems in their ground and excited states, and for studies of energy surfaces for chemical reactions. It is applicable to any stationary state of an atomic or molecular system. It can also be used for open-shell states and far from equilibrium.

In principle, at least, CI is capable of giving accurate solutions to the non-relativistic clamped-nuclei Schrödinger equation of subsection 2.1.5. It has a lot of computational difficulties but its conceptual simplicity and generality make it very appealing for chemists.

In this section we extend what has already been said about configuration interaction. In particular we look at the CI eigenvalue problem. The article by Shavitt in [35, pages 189–275] gives further details of a lot of what is looked at in this section.

6.1.1 The CI Equations Revisited

Let us review the equations given in subsection 4.2.6. In the CI method we optimize the coefficients c_I in

$$\Psi = \sum_I c_I \Psi_I \tag{6.1}$$

by using the Rayleigh-Ritz variation principle. The Ψ_I are configuration state functions (CSF's) and therefore satisfy the right sort of spin and symmetry conditions. The coefficients c_I are chosen so that the expectation value of the energy is minimized. In the CI method we solve the secular equations

$$H\mathbf{c} = E\,S\mathbf{c}$$

where the matrices H and S are given by

$$H_{IJ} = \langle \Psi_I | \mathcal{H}_{\text{elec}} | \Psi_J \rangle$$

$$S_{IJ} = \langle \Psi_I | \Psi_J \rangle$$

The energy E is given by

$$E = \frac{\langle \Psi | \mathcal{H}_{\text{elec}} | \Psi \rangle}{\langle \Psi | \Psi \rangle}$$

and the column vector $\mathbf{c}$ contains the CI coefficients c_I which are used in equation (6.1) to define Ψ.

When the CSF's are orthonormal $S_{IJ} = \delta_{IJ}$ and the secular equations have the simpler form

$$H\mathbf{c} = E\,\mathbf{c}$$

In [7, pages 95–113] Davidson looks at when the ordinary and generalized eigenvalue problems get solved. The generalized one is used for expansions with a few electrons and the S matrix is often ill-conditioned due to near linear dependences in the configuration basis. Orthogonal CI expansions are used for many electron problems and it is this case that we focus on here. The ordinary eigenvalue problems tend to be a lot larger in dimension with matrices of order up to 10^9.

Recall that one of the reasons for performing a CI calculation is to improve the accuracy of the results. We want to reduce the magnitude of the correlation energy

$$E_{\text{corr}} = \mathcal{E}_0 - E$$

where $\mathcal{E}_0$ is the true non-relativistic energy and E is the energy of the CI wave function. The correlation energy is negative because of the variation principle. The correlation energy is usually a small percentage of the total energy of a molecular system, eg. 0.5 % in H_2O [35, page 189]. Chemistry is primarily concerned with small energy differences such as those between different electronic states or between different geometries, eg. the binding energy of H_2O is also 0.5 % of the total energy. Thus energy differences can be seriously affected by the correlation error.

We usually split the energy E into two parts, the SCF energy, and the part of the energy which is due to the correlation in the wave function. We can think of this as the CI correction, and it will not be as large as E_{corr} in magnitude. It can be divided into two parts, from dynamical and non-dynamical correlation.

The Hartree-Fock method can only be used for processes in which the correlation error can be assumed not to vary. This is not true for a lot of problems and CI is often used for such problems.

In the CI method each CSF is already known (unlike with the MCSCF method) and can be expressed as sums of products of one-electron orbitals. They are linear combinations of Slater determinants that are usually made up from orthonormal MO's $\{\psi_i\}$. The CSF's are usually required to satisfy some or all of the boundary and symmetry conditions that the wave function Ψ is required to satisfy. A linear combination of Slater determinants is needed, in the general case, to satisfy the spin and symmetry conditions.

The matrix element H_{IJ} is zero if $|\Psi_I\rangle$ and $|\Psi_J\rangle$ belong to different symmetry types. So symmetry can be used to divide the problem into smaller problems, one for each symmetry type.

If the molecular orbitals $\{\psi_i\}$ are chosen to be orthonormal then the calculation of the matrix elements H_{IJ} will be greatly simplified. If the CSF's are not orthonormal then the calculation of the S_{IJ} is also simplified. If the $\{\psi_i\}$ are orthonormal then

$$H_{IJ} = \langle \Psi_I | \mathcal{H}_{\text{elec}} | \Psi_J \rangle = \sum_{ij} a_{ij}^{IJ}\, h_{ij}^{[\text{MO}]} + \sum_{ijkl} b_{ijkl}^{IJ}\, g_{ijkl}^{[\text{MO}]}$$

where the h_{ij} and g_{ijkl} are in terms of the $\{\psi_i\}$. The coefficients a_{ij}^{IJ} and b_{ijkl}^{IJ} depend on the matching of orbitals and spin couplings between Ψ_I and Ψ_J. We can express the h_{ij} and g_{ijkl} in terms of the basis functions like we did with the SCF method. As with that method spatial symmetry of the molecule can be used to simplify the number of integrals.

6.1.2 The Conventional CI Method

The computational steps for a **conventional CI** calculation are:

(1) Choose a basis set $\{\phi_\mu\}$ and compute the basis-set integrals.

(2) Choose the molecular orbitals $\{\psi_i\}$ (by a SCF or similar calculation), and transform the basis-set integrals to the molecular orbital integrals.

(3) Choose and construct a set of CSF's of the appropriate spin and space symmetry for the state (or states) being considered, and compute the Hamiltonian matrix H in terms of the CSF's.

(4) Compute the lowest eigenvalue(s) and corresponding eigenvector(s) of the matrix H.

Various additional procedures and iterative loops can be used to choose the orbitals and for the selection of the CSF's. They will not be discussed.

The MO integrals $h_{ij}^{[\text{MO}]}$ and $g_{ijkl}^{[\text{MO}]}$ are related to the AO integrals $h_{ij}^{[\text{AO}]}$ and $g_{ijkl}^{[\text{AO}]}$ respectively by

$$h_{ij}^{[\text{MO}]} = \sum_{pq} C_{pi}^*\, C_{qj}\, h_{pq}^{[\text{AO}]} \tag{6.2}$$

$$g_{ijkl}^{[\text{MO}]} = \sum_{pqrs} C_{pi}^*\, C_{qj}\, C_{rk}^*\, C_{sl}\, g_{pqrs}^{[\text{AO}]} \tag{6.3}$$

The matrix C relates the AO's $\{\phi_\mu\}$ to the MO's $\{\psi_i\}$ according to

$$\psi_i = \sum_{\mu=1}^{N_{\text{BF}}} C_{\mu i}\, \phi_\mu \qquad\qquad i = 1, 2, \ldots, N_{\text{BF}}$$

We now consider the propagation of errors. In SCF calculations this is usually not all that serious, with the errors in the basis set integrals being reflected by errors of not much greater magnitude in the final energies. In a SCF wave function the coefficients C_{pi} are $O(1)$ for the occupied orbitals, so there is no amplification of errors in the two transformations for the occupied orbitals. This step is implicit in the construction of the Fock matrix. In CI calculations the virtual orbitals are also used and with a large basis set that is nearly linearly dependent some of the C_{pi} could be one or more orders of magnitude greater than unity. So this means the errors are greatly amplified in equations (6.2) and (6.3), particularly in (6.3). So it is very important to get high accuracy in integral evaluations for large basis-set CI calculations. Sometimes it is necessary to discard some of the orbitals. The magnitude of the largest C_{pi} and smallest c_I need to be looked at when considering the propagation of errors.

6.1.3 The Direct-CI Method

At the end of chapter 4 direct configuration interaction methods were very briefly mentioned. They avoid explicitly constructing the Hamiltonian matrix H. Direct methods can lead to very efficient algorithms for working out the CI coefficients and remove the major storage bottleneck for very large CI calculations.

The first direct-CI method was introduced by Roos in 1972. Direct-CI is looked at by Saunders and van Lenthe in [33], Siegbahn in [7, pages 189–207], and Roos and Siegbahn in [35, pages 277-318]. It has become the basis of recent advances in the calculation of correlated wave functions.

Iterative eigenvalue methods that require only the operator form of matrices are needed for the direct-CI method. The term direct comes from the fact that the wave functions are constructed *directly* without constructing an intermediate Hamiltonian matrix. The matrix-vector multiplication

$$\mathbf{z} = H\,\mathbf{c}$$

where $\mathbf{c}$ contains the trial CI expansion coefficients, is done from the molecular integrals. Each Hamiltonian matrix element is just a linear combination of molecular integrals. A complete transformation of the molecular integrals from an AO to a MO basis set is required.

The task of direct-CI is to construct the wave function efficiently. It is pointed out in [33] that there is no single optimal procedure for carrying out direct-CI calculations. A general code should work in all the possible options and be capable of selecting the best strategy. We will not go into any more detail because the discussion is not important here, and it becomes quite involved.

6.1.4 Basis Sets

With CI calculations more consideration of the choice of basis set is needed than with SCF calculations. This is because the basis set has to provide appropriate **correlation orbitals** as well as the occupied orbitals of the SCF wave function. These orbitals have their greatest amplitude in regions where the occupied SCF orbitals, and therefore the electron density, are mainly concentrated. However they also have additional nodal surfaces that divide these regions in lots of ways. Not even a full-CI calculation can make up for an inadequate basis set in the underlying SCF calculation, and in a full-CI calculation the result depends particularly on the choice of a basis set. All the things in section 3.3 about basis sets also apply to CI wave functions as far as the highly occupied part of the orbital space is concerned. However even here there are some additional considerations.

In large scale conventional-CI calculations the evaluation of the basis set integrals is not such a large fraction of the computational time as in a SCF calculation. Therefore for a CI calculation the consideration of the integral time is not as important. So Slater-type orbitals are more likely to be used as the basis functions.

The basis set has to provide for the nodal surfaces of the correlation orbitals.

Orbital exponent optimization is a lot harder for CI wave functions than SCF ones.

A deep insight into the chemical problem is often necessary because in practice the basis set is far from complete.

6.1.5 Features of the CI Eigenvalue Problem

CI leads to the formation of a large, real, symmetric matrix for which the lowest eigenvalues and corresponding eigenvectors must be found. The CI energy E

and expansion coefficients vector **c** are the appropriate eigenvalue and corresponding eigenvector of the Hamiltonian matrix H. If the CSF's are not orthonormal the generalized eigenvalue equation $H\mathbf{c} = E\,S\mathbf{c}$ has to be solved. We will assume that the CSF's have been orthonormalized.

The following need to be considered when choosing how to solve the CI eigenvalue problem:

- The sparseness of H.
- The size of H.
- The dominance of the main diagonal in H.
- The number of eigenvalues and corresponding eigenvectors that are required.

Generally about $1 - 5\,\%$ of the matrix elements are non-zero and less than 10 eigenvectors are required [25]. The nonzero elements tend to be randomly distributed. Chemists refer to such a matrix as sparse. Mathematicians will call a matrix with very "few" nonzero entries sparse. There are different definitions of the "few". Generally the chemists' matrices are not sparse to a mathematician. The *sparseness* of H reduces the storage. It is useful only if the method that is used preserves the zeros. Standard direct methods do not do this and such methods are only practical for relatively small CI problems. An iterative method that does not modify the original matrix in intermediate stages is most suitable for the CI eigenvalue problem.

The handling of H also needs to be considered and to do this how the matrix elements are accessed needs to be looked at.

The size of H often means that the whole matrix cannot be stored in the central memory of the computer. The dimension of the matrix is usually larger than 10^4 and the non-zero elements are stored on disk. For larger scale calculations, with matrix dimensions greater than 10^5, the matrix may be too large to store on disk and the elements must be recalculated as they are needed. This is so with the direct-CI method. Matrices of dimension 10^6 have been used and even of dimension 10^9 [25].

Many electronic states are described reasonably well by a single CSF and this causes the dominance of the main diagonal. Typically this results in the eigenvector being dominated by one large component since the eigenvectors contain the CI coefficients.

If only one or very few of the lowest eigenvalues and eigenvectors are needed then H does not need to be completely diagonalized. The higher eigenvalues and eigenvectors could be used as approximations to higher excited states, but they are probably not very good. The basis set and CSF's would have been chosen for the lower states.

It should be noted that reasonably good initial guesses are usually available. They can be generated by looking at solutions that are similar to the desired solutions. Several iterations can be saved in most methods if we start with initial guesses that have dot products exceeding about 0.7 with the exact result [5]. However most methods converge fast enough that it is not worth the extra effort of generating more accurate guesses.

Davidson has given a couple of reviews of the eigenvalue problems in quantum chemistry [5] and [7, pages 95–113]. Both talk about CI and Davidson's method.

6.2 Lanczos Methods

We now look at the mathematical background for Davidson's method. We are working up to looking at the Lanczos algorithm. It dates back to 1950 and it flopped then because it is prone to rounding errors. In about 1970 it was shown to be effective for computing some outer eigenpairs. It is a powerful technique for computing a few eigenvalues of a symmetric matrix. It can be described as the natural way to implement the Rayleigh-Ritz procedure on a sequence of Krylov subspaces. We start by looking at what those things are.

The four books Parlett [26], Golub and Van Loan [11], Trefethen and Bau [43], and Saad [32] were used.

Throughout this section we will consider the matrix A which is real and symmetric.

6.2.1 The Rayleigh-Ritz Procedure

The **Rayleigh-Ritz procedure** is a way of computing the *best* set of approximate eigenvectors for a matrix A from a subspace. We will look at the sense in which they are best later, on page 102.

Suppose A is $n \times n$, with eigenvalues λ_i, labelled in increasing order, and corresponding eigenvectors $\mathbf{z}_i$.

A subspace $\mathcal{S}$ is **invariant** under A if $A\,\mathcal{S} \subset \mathcal{S}$. Any invariant subspace has a basis of eigenvectors.

We need to have a way of testing a set of vectors to see if they span an invariant subspace. Let $F = [\,\mathbf{f}_1, \ldots, \mathbf{f}_m\,]$ be an $n \times m$ matrix. If F is invariant then

$$A\,\mathbf{f}_j = \sum_i \mathbf{f}_i\, C_{ij} \qquad\qquad j = 1, \ldots, m$$

Consider F's **residual matrix**, which is defined by

$$R(F) \;\equiv\; \boxed{\quad A \quad}\,\boxed{F} - \boxed{F}\,\boxed{C} = \boxed{0}$$

Suppose F has full rank so we can solve for a unique C

$$C = (F^T F)^{-1} F^T A F$$

If F did not have full rank then there would be many such C's, and we do not need to consider this case.

Suppose Q is an orthonormal basis of span(F) ($=$ span(Q)). The test for invariance is

$$R(Q) \equiv AQ - QH = [0] \qquad\qquad \text{where} \;\; H = Q^T A Q$$

Both C and H represent the restriction of A to span(F). As H is symmetric it is therefore in a sense better.

The **Rayleigh quotient** of a nonzero vector $\mathbf{x}$ is given by

$$r(\mathbf{x}) \equiv \frac{\mathbf{x}^T A \mathbf{x}}{\mathbf{x}^T \mathbf{x}} \tag{6.4}$$

The minimum value of this function is $\lambda_1(A)$, the smallest eigenvalue of A. We can also define the Rayleigh quotient for orthogonal matrices, and H is the Rayleigh quotient of Q. It can also be defined for an orthonormal wave function, which obviously related to the expectation value of the Hamiltonian and the Rayleigh-Ritz variation principle.

Each eigenvector of C, or H, determines an eigenvector of A.

$$\text{If} \quad C\mathbf{y} = \mathbf{y}\lambda \quad \text{then} \quad A(F\mathbf{y}) = (F\mathbf{y})\lambda.$$

$$\text{If} \quad H\mathbf{x} = \mathbf{x}\lambda \quad \text{then} \quad A(Q\mathbf{x}) = (Q\mathbf{x})\lambda.$$

Suppose S is a full rank $n \times m$ matrix whose columns $\mathbf{s}_i$ span $\mathcal{S}^m$ which is an m dimensional subspace of R^n. Usually $\mathcal{S}^m$ is not invariant under A. If it is nearly invariant then it should contain good approximations to some of A's eigenvectors. This is the idea behind the Rayleigh-Ritz procedure.

The **Rayleigh-Ritz procedure** is:

(1) Orthonormalize the columns of S to get an orthogonal $n \times m$ matrix Q.

(2) Form the $m \times m$ matrix H, which is the Rayleigh quotient of Q,

$$H = \rho(Q) = Q^T A Q$$

(3) Compute the p ($\leq m$) eigenpairs of H that are wanted,

$$H\mathbf{g}_i = \mathbf{g}_i \theta_i \qquad i = 1, \ldots, p$$

The θ_i are the **Ritz values**.

(4) Compute the p **Ritz vectors**,

$$\mathbf{y}_i = Q\,\mathbf{g}_i \qquad i = 1, \ldots, p$$

(5) Form the p residual vectors

$$\mathbf{r}_i = r(\mathbf{y}_i) = A\,\mathbf{y}_i - \mathbf{y}_i\theta_i = AQ\,\mathbf{g}_i - \mathbf{y}_i\theta_i$$

and compute $||\mathbf{r}_i||_2$.

The full set $\{(\theta_i, \mathbf{y}_i), i = 1, \ldots, m\}$ is the best set of approximations to eigenpairs of A that can be obtained from $\mathcal{S}^m$ alone.

We calculate the $||\mathbf{r}_i||_2$ because each of the intervals $[\,\theta_i - ||\mathbf{r}_i||_2, \theta_i + ||\mathbf{r}_i||_2\,]$ contains an eigenvalue of A. If some of the intervals overlap then more work needs to be done [26, pages 218–220].

It should be noted that there was no need to know A explicitly in the procedure. All we need is an algorithm that computes $A\mathbf{x}$ from $\mathbf{x}$. We can greatly reduce the computation if A has a special structure, eg. if A is sparse. This is very important with direct-CI methods.

We now look at the sense in which the Ritz pairs are best. There are three related ways of explaining how they are optimal for the given information [26, pages 215–217].

Using the minimax characterization of eigenvalues we can justify why the θ_i are optimal. As we have already said

$$\lambda_1(A) = \min\{\rho(\mathbf{x}) : \mathbf{x} \neq 0\}$$

The higher eigenvalues can be expressed as constrained minimums,

$$\lambda_i(A) = \min\left\{\rho(\mathbf{x}) : \mathbf{x} \neq 0 \text{ and } \mathbf{z}_j^T\mathbf{x} = 0 \text{ for } j < i\right\}$$

where the $\mathbf{z}_j$ are the eigenvectors for the lower eigenvalues λ_j. This depends explicitly on all the previous eigenvectors. The following theorem, which gives the **minimax characterization**, does not have this problem. We will assume that the Rayleigh quotient is only defined for nonzero vectors.

Theorem: *For $i = 1, \ldots, n$,*

$$\lambda_i(A) = \min_{\mathcal{S}^i} \max_{\mathbf{u}\in\mathcal{S}^i} \rho(\mathbf{u}) = \max_{\mathcal{R}^{i-1}} \min_{\mathbf{v}\perp\mathcal{R}^{i-1}} \rho(\mathbf{v})$$

where $\mathcal{S}^i$ and $\mathcal{R}^{i-1}$ are subspaces of R^n.

The proof of this is given in [26, pages 190-191]. The theorem says that since the function ρ is continuous on the unit sphere, it must attain its maximum on each i-dimensional unit sphere, and the minimum of all these maximum values is λ_i. This gives us a mental picture of what it is saying.

Looking at the characterization of λ_i above the natural definition of the best approximation β_i to λ_i from the subspace $\mathcal{S}^m$ is

$$\beta_i \equiv \min_{\mathcal{G}^i\subset\mathcal{S}^m} \max_{\mathbf{g}\in\mathcal{G}^i} \rho(\mathbf{g})$$

We have the following result.

Theorem:

$$\beta_i = \lambda_i(H) = \theta_i \qquad i = 1, \ldots, m \qquad \text{where } H = Q^T AQ$$

So the θ_i are optimal in terms of the subspace minimax characterization.

The second way they are optimal concerns Q. We can define a residual matrix for any $m \times m$ B

$$R(B) \equiv AQ - QB \tag{6.5}$$

The following theorem states that the minimizing property of the Rayleigh quotient is inherited by $H = Q^T AQ = \rho(Q)$.

Theorem: *Given an $n \times m$ orthonormal matrix Q*

$$\|R(H)\|_2 \leq \|R(B)\|_2$$

for all $m \times m$ B.

If S is any orthonormal basis in $\mathcal{S}^m$ and Δ is any diagonal matrix, the pairs $\{(\delta_i, \mathrm{s}_i), i = 1, \ldots, m\}$ are rival eigenpair approximations. From the theorem above we can prove that $\|A\mathrm{s} - \mathrm{s}\Delta\|$ is minimized over S and Δ when and only when

$$\mathrm{s}_i = \mathbf{y}_i \qquad \delta_i = \theta_i \qquad i = 1, \ldots, m$$

The final way that the Rayleigh-Ritz approximations are optimal is to do with projections. $\mathcal{S}^m$ is not invariant so consider A's projection onto $\mathcal{S}^m$ rather than the restriction of A to $\mathcal{S}^m$. We will not go into the details of this, and just state the following: *The (θ_i, y_i), $i = 1, \ldots, m$ are the eigenpairs for A's projection onto $\mathcal{S}^m$.*

There are a couple of ways in which the Rayleigh-Ritz approximations are not optimal. Generally no $\mathbf{y}_i$ is the closest unit vector in $\mathcal{S}^m$ to any eigenvector of A. Also the error bound $\|A\mathrm{v} - \mathrm{v}\rho(\mathrm{v})\|/\|\mathrm{v}\|$ is not minimized in $\mathcal{S}^m$ by any of the Ritz vectors for $m > 1$. In summary, by getting *collective* optimality for m pairs, the Rayleigh-Ritz approximation usually gives up individual optimality for any pair.

6.2.2 Krylov Subspaces

Krylov subspaces are important in the theory of various eigenvalue methods. Suppose $\mathbf{f}$ is a nonzero vector. The **Krylov matrices** $K^m(\mathbf{f})$ are given by

$$K^m(\mathbf{f}) \equiv \left[\mathbf{f}, A\mathbf{f}, \ldots, A^{m-1}\mathbf{f}\right] \tag{6.6}$$

The **Krylov subspace** $\mathcal{K}^m(\mathbf{f})$ is defined by

$$\mathcal{K}^m(\mathbf{f}) \equiv \operatorname{span} K^m(\mathbf{f}) \tag{6.7}$$

Usually $\mathcal{K}^m(\mathbf{f})$ is of dimension m unless $\mathbf{f}$ is related to A or $m > n$.

If $\mathbf{f}$ is the starting vector in the power method then the columns of $K^m(\mathbf{f})$ get computed one by one. The Rayleigh-Ritz procedure involves more storage and computation than the power method. But it gives better approximations and it turns out to be cost effective in general. For example, the Rayleigh-Ritz approximations are exact for $\mathcal{K}^n(\mathbf{f})$, whereas in principal the power method takes an infinite number of iterations. On the other hand there is a possibility that $\mathbf{f}$ is orthogonal to the eigenspace, and therefore $\mathcal{K}^m(\mathbf{f})$ is orthogonal to the eigenspace too for all m.

For theoretical work the natural basis for $\mathcal{K}^m(\mathbf{f})$ is the columns of $K^m(\mathbf{f})$. For practical work it is the orthonormal basis $Q_m \equiv [\,\mathbf{q}_1, \ldots, \mathbf{q}_m\,]$ which comes from applying the Gram-Schmidt process to the columns of $K^m(\mathbf{f})$ in the order $\mathbf{f}, A\mathbf{f}, \ldots$. This is the **Lanczos basis.**

6.2.3 The Lanczos Algorithm

The Lanczos method can be used to solve certain large, sparse, symmetric eigenproblems $A\mathbf{x} = \lambda\mathbf{x}$. It involves partial tridiagonalizations of A. In contrast with the Householder approach no intermediate full sub-matrices are generated. Information about A's extremal eigenvalues tends to emerge long before the tridiagonalization is complete. So Lanczos is particularly good when a few of A's largest or smallest eigenvalues are desired.

Roundoff error makes the Lanczos method hard to use. The central problem is a loss of orthogonality among the Lanczos vectors. This makes it harder to tell when to terminate the algorithm, and it complicates the relationship between A's eigenvalues and those of the tridiagonal matrices T_k. In the three-term recurrence relation that is used the orthogonality is expected to arise automatically. But with rounding error the orthogonality is lost after a few iterations. This could be corrected by using the Gram-Schmidt algorithm at each iteration.

Suppose $A \in \mathrm{R}^{n\times n}$ is large, sparse and symmetric. Assume that a few of its largest and/or smallest eigenvalues are needed. The method generates a sequence of tridiagonal matrices T_k with the property that the extremal eigenvalues of $T_k \in \mathrm{R}^{k\times k}$ are progressively better estimates of A's extremal eigenvalues.

Let λ_i be the ith largest eigenvalue. Note that we have ordered the eigenvalues the opposite way in the Schrödinger equation.

We derive the Lanczos algorithm by considering the optimization of the Rayleigh quotient

$$r(\mathbf{x}) = \frac{\mathbf{x}^T A \mathbf{x}}{\mathbf{x}^T \mathbf{x}} \qquad \mathbf{x} \neq \mathbf{0}$$

The maximum and minimum values of $r(\mathbf{x})$ are $\lambda_1(A)$ and $\lambda_n(A)$ respectively.

Let $\{\mathbf{q}_i\} \subseteq \mathrm{R}^n$ be orthonormal vectors and $Q_k = [\,\mathbf{q}_1, \ldots, \mathbf{q}_k\,]$. Define

$$M_k = \lambda_1(Q_k^T A Q_k) = \max_{\mathbf{y}\neq 0} \left(\frac{\mathbf{y}^T (Q_k^T A Q_k) \mathbf{y}}{\mathbf{y}^T \mathbf{y}} \right) = \max_{\|\mathbf{y}\|_2 = 1} \left(r(Q_k \mathbf{y}) \right) \leq \lambda_1(A)$$

and

$$m_k = \lambda_k(Q_k^T A Q_k) = \min_{\mathbf{y}\neq 0} \left(\frac{\mathbf{y}^T (Q_k^T A Q_k) \mathbf{y}}{\mathbf{y}^T \mathbf{y}} \right) = \min_{\|\mathbf{y}\|_2 = 1} \left(r(Q_k \mathbf{y}) \right) \geq \lambda_n(A)$$

We can get the Lanczos algorithm by considering how to generate the $\mathbf{q}_k$ so that M_k and m_k are increasingly better approximations. It is easy to show that this can be achieved if

$$\mathrm{span}\{\mathbf{q}_1, \ldots, \mathbf{q}_k\} = \mathrm{span}\{\mathbf{q}_1, A\mathbf{q}_1, \ldots, A^{k-1}\mathbf{q}_1\} \tag{6.8}$$

To get this the fact that $r(\mathbf{x})$ increases most rapidly in the direction of the gradient $\nabla r(\mathbf{x})$ is used. Let $\mathbf{u}_k \in \mathrm{span}\{\mathbf{q}_1, \ldots, \mathbf{q}_k\}$ be such that $M_k = r(\mathbf{u}_k)$. The gradient is

$$\nabla r(\mathbf{x}) = \frac{2}{\mathbf{x}^T \mathbf{x}} (A\mathbf{x} - r(\mathbf{x})\mathbf{x})$$

We can make sure that $M_{k+1} > M_k$ if $\mathbf{q}_{k+1}$ is determined so that

$$\nabla r(\mathbf{u}_k) \in \mathrm{span}\{\mathbf{q}_1, \ldots, \mathbf{q}_{k+1}\}$$

assuming that $\nabla r(\mathbf{u}_k) \neq \mathbf{0}$. Similarly if $\mathbf{v}_k \in \mathrm{span}\{\mathbf{q}_1, \ldots, \mathbf{q}_k\}$, is such that $m_k = r(\mathbf{v}_k)$, then we want

$$\nabla r(\mathbf{v}_k) \in \mathrm{span}\{\mathbf{q}_1, \ldots, \mathbf{q}_{k+1}\}$$

since $r(\mathbf{x})$ decreases most rapidly in the direction of $-\nabla r(\mathbf{x})$. Now since $\nabla r(\mathbf{x}) \in \text{span}\{\mathbf{x}, A\mathbf{x}\}$ we can satisfy the $\nabla r(\mathbf{u}_k)$ and $\nabla r(\mathbf{v}_k)$ conditions if (6.8) holds and we choose $\mathbf{q}_{k+1}$ so that

$$\text{span}\{\mathbf{q}_1, \ldots, \mathbf{q}_{k+1}\} = \text{span}\{\mathbf{q}_1, A\mathbf{q}_1, \ldots, A^{k-1}\mathbf{q}_1, A^k\mathbf{q}_1\}$$

If $\mathbf{x}$ is an element of the space given by (6.8) then $\nabla r(\mathbf{x})$ will be in the above space.

So the problem is to compute an orthonormal basis for the Krylov subspace $\mathcal{K}^k(A, \mathbf{q}_1)$.

To get the Lanczos method we consider tridiagonalizing A with a particular matrix Q. If $Q^TAQ = T$ is tridiagonal with $Q\mathbf{e}_1 = \mathbf{q}_1$, where $\mathbf{e}_1$ is the first column of the identity, then

$$K^n(A, \mathbf{q}_1) = Q[\,\mathbf{e}_1, T\mathbf{e}_1, T^2\mathbf{e}_1, \ldots, T^{n-1}\mathbf{e}_1\,]$$

is the QR factorization of $K^n(A, \mathbf{q}_1)$. So $\mathbf{q}_k$ can be obtained by tridiagonalizing A with an orthogonal matrix whose first column is $\mathbf{q}_1$. Householder tridiagonalization is impractical if A is large and sparse because the sparseness is lost.

By setting $Q = [\,\mathbf{q}_1, \ldots, \mathbf{q}_n\,]$ and

$$T = \begin{bmatrix} \alpha_1 & \beta_1 & \cdots & \cdots & 0 \\ \beta_1 & \alpha_2 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \beta_{n-1} \\ 0 & \cdots & \cdots & \beta_{n-1} & \alpha_n \end{bmatrix}$$

and equating the columns in $AQ = QT$, we get

$$A\mathbf{q}_k = \beta_{k-1}\mathbf{q}_{k-1} + \alpha_k\mathbf{q}_k + \beta_k\mathbf{q}_{k+1} \qquad k = 1, 2, \ldots, n-1 \tag{6.9}$$

where $\beta_0\mathbf{q}_0 \equiv \mathbf{0}$. The orthonormality of the $\mathbf{q}_i$ implies $\alpha_k = \mathbf{q}_k^TA\mathbf{q}_k$ and $\beta_k = \mathbf{q}_{k+1}^TA\mathbf{q}_k$. If $\mathbf{r}_k = (A - \alpha_kI)\mathbf{q}_k - \beta_{k-1}\mathbf{q}_{k-1}$ is nonzero then $\mathbf{q}_{k+1} = \beta_k^{-1}\mathbf{r}_k$ where $\beta_k = \pm\|\mathbf{r}_k\|_2$. If $\mathbf{r}_k = \mathbf{0}$ then the iterations stop.

The **Lanczos iterative procedure** is:

(1) Choose a unit vector $\mathbf{q}_1$ and set the initial values:

$$\mathbf{r}_0 = \mathbf{q}_1; \quad \beta_0 = 1; \quad \mathbf{q}_0 = \mathbf{0}; \quad k = 0$$

(2) Put $\mathbf{q}_{k+1} = \beta_k^{-1}\mathbf{r}_k$.

(3) Set $k = k + 1$.

(4) Compute $\alpha_k = \mathbf{q}_k^T A \mathbf{q}_k$.

(5) Construct $\mathbf{r}_k = (A - \alpha_k I)\mathbf{q}_k - \beta_{k-1}\mathbf{q}_{k-1}$.

(6) Compute $\beta_k = \|\mathbf{r}_k\|_2$.

(7) If $\beta_k \neq 0$ go back to step (2).

β_k is chosen to be positive without loss of generality. The $\mathbf{q}_k$ are called the **Lanczos vectors**. At each step the subspace dimension grows by one. This is not the way the algorithm is implemented in practice because of the rounding error problem that has already been mentioned. Practical ideas for the Lanczos algorithm are discussed in [11, pages 480–488].

The following two theorems are taken from [11, pages 474-475].

Theorem: *Let the matrix $A \in \mathrm{R}^{n\times n}$ be symmetric and assume $\mathbf{q}_1 \in \mathrm{R}^n$ is such that $\|\mathbf{q}_1\|_2 = 1$. Then the Lanczos method runs until $k = m$, where $m =$ rank $(K(A, \mathbf{q}_1, n))$. For $k = 1, 2, \ldots, m$*

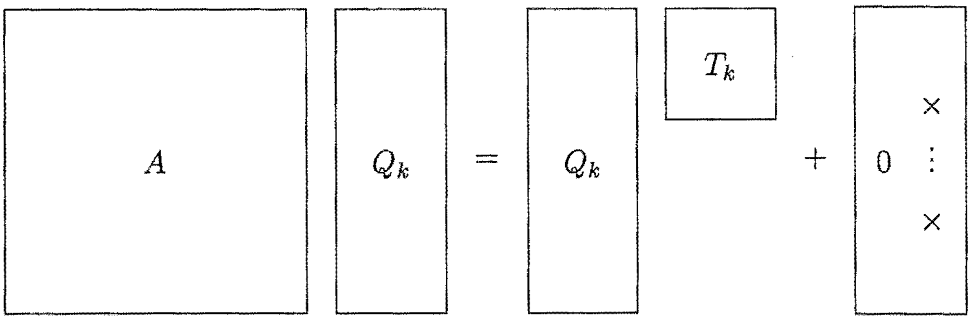
$$AQ_k = Q_k T_k + \mathbf{r}_k \mathbf{e}_k^T \tag{6.10}$$

where

$$T_k = Q_k^T A Q_k = \begin{bmatrix} \alpha_1 & \beta_1 & \cdots & \cdots & 0 \\ \beta_1 & \alpha_2 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \beta_{k-1} \\ 0 & \cdots & \cdots & \beta_{k-1} & \alpha_k \end{bmatrix}$$

and Q_k has orthonormal columns and they span $\mathcal{K}^k(A, \mathbf{q}_1)$.

For each $k < m$ equation (6.10) looks like



where the last column on the right hand side is

$$\mathbf{r}_j \equiv \beta_j \mathbf{q}_{j+1}$$

If we get $\beta_k = 0$ then we have an exact invariant subspace. In this case $AQ_k = Q_k T_k$ and span $Q_k = \mathcal{K}^k(A, \mathbf{q}_1)$ is the smallest invariant subspace containing $\mathbf{q}_1$. In practice a zero or small value of β_k is rare.

The following helps explain why the extremal eigenvalues of T_k are good approximations to those of A.

Theorem: *Suppose we have done k steps of the Lanczos algorithm. Let* $S_k^T T_k S_k = diag(\theta_1, \ldots, \theta_k)$ *be the Schur decomposition of the tridiagonal matrix* T_k*. If*

$$Y_k = [\,\mathbf{y}_1, \ldots, \mathbf{y}_k\,] = Q_k S_k \in \mathrm{R}^{n\times k},$$

then for $i = 1, 2, \ldots, k$ *we have*

$$||A\mathbf{y}_i - \theta_i \mathbf{y}_i||_2 = |\beta_k||s_{ki}|,$$

where the s_{ki} *are the elements of* S*.*

The $(\theta_i, \mathbf{y}_i)$ are Ritz pairs for the subspace span $\{\mathbf{q}_1, \ldots, \mathbf{q}_k\}$.

We can derive the Lanczos procedure by considering Rayleigh-Ritz applied to a sequence of Krylov subspaces with the Lanczos basis. The cost of Rayleigh-Ritz is reduced a lot when a sequence of Krylov subspaces is used. We find that a lot of the things that need to be computed are already on hand. This is the derivation used by Parlett [26].

6.2.4 The Conjugate Gradient Connection

The Lanczos iteration can be used to solve large sparse linear equations and least squares problems.

We now briefly look at how the **conjugate gradient method** can be derived from the Lanczos method. It is the "original" Krylov subspace iteration. It solves symmetric positive definite systems of equations very quickly when the eigenvalues are well separated.

Suppose A is $n \times n$, symmetric and positive definite, and $\mathbf{b} \in \mathrm{R}^n$. Consider the functional $\phi(\mathbf{x})$ given by

$$\phi(\mathbf{x}) \equiv \tfrac{1}{2}\mathbf{x}^T A\mathbf{x} - \mathbf{x}^T\mathbf{b} \tag{6.11}$$

Since $\nabla\phi(\mathbf{x}) = A\mathbf{x} - \mathbf{b}$ the unique minimizer of ϕ is given by $\mathbf{x} = A^{-1}\mathbf{b}$. So an approximate minimizer is an approximate solution to $A\mathbf{x} = \mathbf{b}$.

Let $\mathbf{x}_0 \in \mathrm{R}^n$ be a starting guess. We want a sequence $\{\mathbf{x}_k\}$ that converges to the solution $\mathbf{x}$. We can do this by generating a sequence of orthonormal vectors $\{\mathbf{q}_k\}$ and choosing $\mathbf{x}_k$ to minimize ϕ over

$$\mathbf{x}_0 + \mathrm{span}\,\{\,\mathbf{q}_1, \ldots, \mathbf{q}_k\,\} = \{\,\mathbf{x}_0 + y_1\mathbf{q}_1 + y_2\mathbf{q}_2 + \cdots + y_k\mathbf{q}_k : \; y_i \in \mathrm{R}\,\}$$

If $Q_k = [\,\mathbf{q}_1, \ldots, \mathbf{q}_k\,]$ we want $\mathbf{y} \in \mathrm{R}^k$ such that

$$\phi(\mathbf{x}_0 + Q_k\mathbf{y}) = \tfrac{1}{2}\mathbf{y}^T\left(Q_k^T A Q_k\right)\mathbf{y} - \mathbf{y}^T Q_k^T(\mathbf{b} - A\mathbf{x}_0) + \phi(\mathbf{x}_0)$$

is minimized. This means, from the gradient, that

$$\mathbf{x}_k = \mathbf{x}_0 + Q_k\mathbf{y}_k \tag{6.12}$$

where $\mathbf{y}_k$ is the solution of

$$\left(Q_k^T A Q_k\right)\mathbf{y}_k = Q_k^T(\mathbf{b} - A\mathbf{x}_0) \tag{6.13}$$

When $k = n$, $A\mathbf{x}_n = \mathbf{b}$ because we are minimizing over the whole of R^n. When A is large and sparse equation (6.13) needs to be easy to solve, and $\mathbf{x}_k$ needs to be computed without referring explicitly to $\mathbf{q}_1, \ldots, \mathbf{q}_k$. We can do this if the $\mathbf{q}_k$ are the Lanczos vectors. We can use the LDL^T factorization of T_k,

$$T_k = Q_k^T A Q_k = L_k D_k L_k^T$$

Following along these lines we can get the method of conjugate gradients, and the algorithm is given in [11, page 493].

The use of the conjugate gradient method in molecular electronic structure calculations is looked at by Wormer, Visser and Paldus in [47].

6.2.5 Preconditioning

We now look at the idea of preconditioning, which is used in Davidson's method. It basically involves transforming the problem so that the matrix properties are improved.

Consider the $n \times n$ nonsingular system

$$B\mathbf{x} = \mathbf{c} \tag{6.14}$$

If M is any nonsingular $n \times n$ matrix then

$$M^{-1}B\mathbf{x} = M^{-1}\mathbf{c} \tag{6.15}$$

has the same solution as (6.14). When we solve (6.15) iteratively the convergence depends on properties of $M^{-1}B$ rather than those of B. M is known as a **(left) preconditioner**, and if it is well chosen convergence can be speeded up. Obviously we need to be able to compute $M^{-1}B$ efficiently. By multiplying (6.14) by M^{-1} rather than M we are stressing the fact that M has to be nonsingular.

We want M to be somewhere between the two extremes $M{=}B$ and $M{=}I$. We could use $M{=}\mathrm{diag}\,(B)$ say, as long as it is not singular. The following rule of thumb is adequate [43, page 314]. *A preconditioner M is good if $M^{-1}B$ is not too far from normal and its eigenvalues are clustered.*

Preconditioners can also be used effectively for eigenvalue problems. The Davidson method uses a kind of diagonal preconditioner.

A preconditioner cannot be directly applied to an eigenvalue problem $A\mathbf{x} = \lambda\mathbf{x}$ like it was in equation (6.15). If this is done the problem becomes a generalized eigenvalue problem and this will not be easier to solve. If only A is multiplied by the preconditioner the eigenpairs are changed. So we need to transform the problem in such a way that the eigenpairs for the original problem can be easily obtained.

For eigenvalue problems the best known preconditioner is the **shift-and-invert preconditioner**. Suppose we are looking for the matrix A's eigenvalues. If the shift σ is well chosen the shifted and inverted matrix

$$C = (A - \sigma I)^{-1}$$

will have a "better" spectrum than the original matrix A. If the spectrum is more separated convergence will be faster. The eigenvectors of A and C are identical and it is easy to calculate A's eigenvalues from those of C. The term preconditioner is appropriate because the condition of the matrix is improved. If the eigenvalues are more separated around the desired eigenvalue then the corresponding eigenvector is likely to be better conditioned.

In the next section we look at Davidson's method. The success of the method on some types of eigenvalue problems shows the potential power of diagonal preconditioning.

6.3 Davidson's Method

The original formulation of Davidson's method [6] can be viewed as a cheap version of shift-and-invert in which the linear system is solved (very) inaccurately [32, page 279]. However quantum chemists find the method very useful for solving the problems they are faced with. The method is well-known by chemists and physicists, but not as well-known by numerical analysts. One reason for this could be the lack of theory about the method.

As has already been mentioned Davidson's (generalized) method is a preconditioned version of the Lanczos method. It is a generalization of the Lanczos algorithm because it includes that algorithm. Davidson's method, like the Lanczos method, uses the Rayleigh-Ritz procedure but on what is usually a non-Krylov subspace. Another possible reason for why Davidson's (generalized) method is not well-known by numerical analysts is that it is a very expensive way of implementing the Lanczos algorithm. Knowing this is rather off-putting.

It is worth noting that Davidson's method was originally proposed as a modification of relaxation methods and not as a generalization of the Lanczos method.

We first look at a generalized version of Davidson's method which seems to date back to 1986. Then we consider the original version of the method. Finally the original method's derivation and convergence is looked at.

6.3.1 The Generalized Davidson Method

The generalized method we are about to look at is referred to as Davidson's method in some mathematical literature, for example [32, pages 272–273].

The basic idea is to generate a set of orthogonal vectors and project onto them. At each step the residual vector $\mathbf{r}$ for the current approximation $\{\theta, \mathbf{g}\}$ to the eigenpair is computed. The vector $\mathbf{r}$ is then multiplied by $(M - \theta I)^{-1}$ where M is some preconditioning matrix. In the original algorithm M was the diagonal of A.

The procedure below is for computing the largest eigenvalue of A. It can also be formulated for the smallest eigenvalue. The notation that was used in the Rayleigh-Ritz procedure on page 101 gets used again. A criterion for the convergence of the residual vector is assumed to be available, and we will discuss the choice of preconditioner M_j after the method is given.

The **Generalized Davidson method** can be expressed as:

(1) *Initialize*: Choose an initial unit vector $\mathbf{q}_1$ and a value of the restart parameter m.

(2) Until convergence iterate step (3).

(3) *Inner Loop*: For $j = 1, 2, \ldots, m$

- Compute $\mathbf{w} = A\mathbf{q}_j$.
- Compute $Q_j^T\mathbf{w}$, which is the last column of $H_j = Q_j^T A Q_j$.
- Compute the largest eigenpair $\{\theta, \mathbf{g}\}$ of H_j.
- Compute the Ritz vector $\mathbf{y} = Q_j\mathbf{g}$.
- Compute the residual vector $\mathbf{r} = A\mathbf{y} - \theta\mathbf{y}$.
- Test to see if convergence is reached.
- Compute $\mathbf{p} = M_j\mathbf{r}$ (skip this step when $j = m$).
- Orthogonalize $\mathbf{p}$ against Q_j using the Modified Gram-Schmidt procedure. Call the result $\mathbf{q}_{j+1}$ (skip this step when $j = m$).

(4) *Restart the procedure*: Set $\mathbf{q}_1 = \mathbf{y}$ and go back to step (3).

The preconditioning matrix M_j is usually an approximation of $(A - \lambda I)^{-1}$. The simplest and most common preconditioner is the original one,

$$M_j = (D - \theta\, I)^{-1} \qquad \text{where} \qquad D = \operatorname{diag}(A) \tag{6.16}$$

This can only be an effective choice when A is almost diagonal which is when the matrix of (normalized) eigenvectors is almost the identity. This helps explain why the algorithm is good for the CI eigenvalue problem. This choice does not work if A is actually diagonal. This will be looked at more in subsection 6.3.3. It is important to note that we do not need to use such a simple preconditioner.

If we have

$$M_j = I$$

then the vectors $\mathbf{v}_j$ produced by Davidson's method are the same as those produced by the Lanczos algorithm on page 106. The matrix H_j corresponds to the T_k in the Lanczos algorithm. In the Lanczos procedure the orthogonality is enforced by the recurrence relation. This will give the same result as Davidson's method in exact arithmetic at least. This difference makes the algorithms look different. Davidson's method is a very costly way of implementing the Lanczos algorithm. This is because it does not use the fact that the cost of Rayleigh-Ritz can be greatly reduced when it is applied to a sequence of Krylov subspaces.

The algorithm could be started with more than one vector, and more than one vector could be kept when it is restarted. There are lots of different possibilities and the algorithm gets stated in different ways.

It is possible to modify the generalized Davidson method to force global convergence to a particular eigenvalue [24, pages 823–824]. If we want the eigenvalue of A closest to σ, then using $(M - \sigma\, I)^{-1}$ as the preconditioner instead of $(M - \theta\, I)^{-1}$ until θ has started to converge improves the global convergence.

There is little known about the convergence of Davidson's method. There is a general convergence result due to Sadkane [32, pages 273–275]. It also applies when more than one eigenvalue is computed.

6.3.2 Davidson's (Original) Method

We now look at the original version of Davidson's method [6] which is commonly used by chemists. It uses the simple preconditioner of equation (6.16).

We start by restating Davidson's method so that it is given with this preconditioner. A formulation similar to that used by Morgan and Scott [24, page 818] is used here. The Rayleigh-Ritz procedure is not built in. It is started with k starting vectors.

Davidson's method is:

(1) *Initialize*: Choose an initial trial space $P_k = \text{span}\{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_k\}$, and compute $\{\mathbf{y}_k, \theta_k\}$ the best approximation to the eigenpair of interest using the Rayleigh-Ritz procedure. Compute the residual vector $\mathbf{r}_k = A\mathbf{y}_k - \theta_k \mathbf{y}_k$. Also choose a value of the restart parameter m.

(2) Until convergence iterate step (3).

(3) *Inner Loop*: For $j = k+1,\ k+2, \ldots,\ k+m$

- Compute $\mathbf{p}_j = (D - \theta_{j-1} I)^{-1} \mathbf{r}_{j-1}$.
- Set $P_j = \text{span}\{P_{j-1}, \mathbf{p}_j\}$.
- Compute $\{\mathbf{y}_j, \theta_j\}$ from P_j by using the Rayleigh-Ritz procedure, which is described on page 101.
- Compute the residual vector $\mathbf{r}_j = (A - \theta_j I)\mathbf{y}_j$.
- Test to see if convergence is reached. This is measured by the norm of the residual vector.

(4) *Restart the procedure*: Choose which vector(s) are going to be kept, and relabel so the vectors span P_k (the value of k might change). Work out the appropriate $\mathbf{y}_k$, θ_k and $\mathbf{r}_k$. Go back to step (3).

The method has been stated with the norm of the residual being small as the convergence criterion. Another test is looking at when the weight of the latest addition to the P_j subspace drops below some threshold. In applications with CI matrices a typical threshold for both tests is 10^{-4} [25].

This algorithm is a lot more expensive per step than the Lanczos method because a full Gram-Schmidt process is needed to compute an orthogonal basis for the space. Also a full reduced matrix is generated by the Rayleigh-Ritz procedure, compared with a tridiagonal one in the Lanczos procedure. Even though it is more expensive it gives better results than Lanczos in CI calculations. It is important to note that Davidson's method usually takes between 10 and 20 iterations to reach convergence when it is used for CI problems.

The new vector at step j is just

$$\mathbf{p}_j = (D - \theta_{j-1}I)^{-1}(A - \theta_{j-1}I)\mathbf{y}_{j-1}$$

This vector is called the **Davidson vector**. As we will see below, it is the correction that is obtained by one step of the Jacobi method for solving the linear system

$$(A - \theta_j I)\,\mathbf{x} = \mathbf{0}$$

with $\mathbf{y}_j$ as the initial guess for $\mathbf{x}$. The articles [25] and [44] both "derive" Davidson's method in this way. This looks strange because $(A - \theta_j I)$ is not singular (except when θ_j is exactly equal to an eigenvalue of A) and the only solution to the system is $\mathbf{x} = \mathbf{0}$. When θ_j is close to the eigenvalue of interest the system will hopefully have a solution that is close to an eigenvector. We obviously need to look at how well-conditioned the problem is to see if this is true. However the Jacobi method has poor convergence and we are only doing one iteration. Recall that in the Jacobi method the matrix is split into $S - T$ where S is the diagonal of the matrix of interest, so it is $(D - \theta_j I)$ in this case, and $T = S - (A - \theta_j I)$. The new trial vector is

$$S^{-1}T\,\mathbf{y}_j = S^{-1}\left(S - (A - \theta_j I)\right)\mathbf{y}_j = \mathbf{y}_j - \left(D - \theta_j I\right)^{-1}\left(A - \theta_j I\right)\mathbf{y}_j$$

The correction is $-(D - \theta_j I)^{-1}(A - \theta_j I)\mathbf{y}_j$. The dropping of the negative sign is unimportant as the vector is added to a space.

We next look at a different explanation of why the form of the new trial vector is chosen. It gives more insight into the convergence of the algorithm.

6.3.3 The Convergence of the Original Algorithm

Let $\{\mathbf{y}, \theta\}$ be the current approximation to the desired eigenpair, with $||\mathbf{y}||_2 = 1$. We have dropped the subscript j. For a given coordinate i, the best approximation that can be made by perturbing $\mathbf{y}$'s ith component can be determined by the Rayleigh-Ritz procedure. Let X be the $n \times 2$ matrix

$$X = [\,\mathbf{y}, \mathbf{e}_i\,]$$

As we will show below the best approximation to the eigenpairs of A from the space span$\{\mathbf{y}, \mathbf{e}_i\}$ is calculated from the 2×2 generalized eigenvalue problem

$$(X^TAX)\mathbf{s} = \alpha(X^TX)\mathbf{s} \tag{6.17}$$

We can write this as

$$(H_i - \alpha W_i)\mathbf{s} = \mathbf{0}$$

where

$$H_i = X^TAX = \begin{bmatrix} \mathbf{y}^TA\mathbf{y} & \mathbf{y}^TA\mathbf{e}_i \\ \mathbf{e}_i^TA\mathbf{y} & \mathbf{e}_i^TA\mathbf{e}_i \end{bmatrix} = \begin{bmatrix} \theta & (A\mathbf{y})_i \\ (A\mathbf{y})_i & a_{ii} \end{bmatrix}$$

and

$$W_i = X^TX = \begin{bmatrix} \mathbf{y}^T\mathbf{y} & \mathbf{y}^T\mathbf{e}_i \\ \mathbf{e}_i^T\mathbf{y} & \mathbf{e}_i^T\mathbf{e}_i \end{bmatrix} = \begin{bmatrix} 1 & y_i \\ y_i & 1 \end{bmatrix}$$

y_i is the ith component of $\mathbf{y}$.

We get a generalized eigenvalue problem because X is not an orthogonal matrix. We now look at why solving this problem gives the best approximations by considering the Rayleigh-Ritz procedure on page 101. We can orthonormalize the columns X to give an orthogonal matrix Q. Put $c = \|\mathbf{y} - y_i\mathbf{e}_i\|_2 = \sqrt{\|\mathbf{y}\|_2^2 - y_i^2}$ and let

$$Q = [\,c^{-1}(\mathbf{y} - y_i\mathbf{e}_i)\,,\ \mathbf{e}_i\,]$$

so that

$$Q = XV \qquad \text{where} \qquad V = \begin{bmatrix} c^{-1} & 0 \\ -c^{-1}y_i & 1 \end{bmatrix}$$

Note that $Q^TQ = I$. The best eigenpairs from the subspace span X, come from solving

$$(Q^TAQ)\mathbf{g} = \alpha(Q^TQ)\mathbf{g}$$

We can express this as

$$V^T(X^TAX)V\mathbf{g} = \alpha V^T(X^TX)V\mathbf{g}$$

By multiplying through by V^{-T} and putting $\mathbf{s} = V\mathbf{g}$ we get the generalized eigenvalue equation (6.17).

For the system (6.17) α is the Ritz value and $\mathbf{z}=Q\mathbf{g}=X\mathbf{s}$ is the corresponding Ritz vector.

Now suppose we are close to convergence. The residual of the current approximation $\mathbf{y}$ will be small and one of the α's will be near θ. The corresponding eigenvector can be approximated by looking at the matrix

$$H_i - \theta W_i = \begin{bmatrix} 0 & (A\mathbf{y}-\theta\mathbf{y})_i \\ (A\mathbf{y}-\theta\mathbf{y})_i & a_{ii}-\theta \end{bmatrix} = \begin{bmatrix} 0 & r_i \\ r_i & d_i-\theta \end{bmatrix}$$

where r_i is the ith component of the residual vector, and d_i is the ith diagonal element of A. The eigenvector is approximately

$$\mathbf{s} \simeq \begin{pmatrix} 1 \\ -(d_i-\theta)^{-1}r_i \end{pmatrix}$$

and so the Ritz vector is

$$\mathbf{z} = X\mathbf{s} \simeq \mathbf{y} - (d_i-\theta)^{-1}r_i\mathbf{e}_i$$

Davidson's method lumps all of these perturbations into one vector

$$\mathbf{y} - (D-\theta I)^{-1}\mathbf{r}$$

This composite vector is added to the space. Actually the second term in the vector above is what is added and the minus sign is dropped, to give the Davidson vector,

$$\mathbf{p} = (D-\theta I)^{-1}\mathbf{r} = (D-\theta I)^{-1}(A-\theta I)\mathbf{y}$$

As has already been mentioned this is change is not important because the space is the same, the bases are all that is different.

The Lanczos algorithm tends to have better global convergence than Davidson's method. In the above we only showed why the algorithm works near the solution.

By looking at the appropriate power series in the components of the Davidson vector $\mathbf{p}$ it is possible to show that Davidson's method does implement the correct first order perturbation correction provided that the components of $\mathbf{p}$ have modulus less than unity [24, pages 819–820]. This is a way of showing more formally that the method works near the solution.

Another way of understanding how Davidson's method behaves is to look at the operator $N(\theta) = (D - \theta)^{-1}(A - \theta I)$. Each new trial vector is just $N(\theta)$ times some vector already in the space. The scalar θ converges to the desired eigenvalue λ, and the properties of $N(\theta)$ for θ near λ help a lot in explaining the properties of the algorithm. See [24, pages 821-822] for further discussion of this operator and the convergence of Davidson's method.

We now consider how different properties of the diagonal of A affect the convergence. We have $r_i \to 0$ and therefore the components of $\mathbf{p}$ will also go to zero unless λ is equal to some diagonal element of A. If this is the case Davidson's method may not perform well. It can almost be said that Davidson's method works better if the matrix is more diagonally dominant which would help explain why the method works so well for CI problems. However if the diagonal is constant Davidson's method is equivalent to Lanczos no matter how diagonally dominant the matrix is. If the diagonal is almost constant then Davidson's method may be slightly faster than the Lanczos algorithm, but it probably will not be worth the extra cost. If A is a diagonal matrix then the new trial vector will just be $\mathbf{y}$, the basis becomes linearly dependent, and so Davidson's method fails completely in this case. In quantum chemistry the matrix A has a large variation in size among the diagonal elements and is not actually a diagonal matrix [5]. Therefore the things we have mentioned should not cause problems in practice for CI calculations.

Since we have only shown that Davidson's method converges locally it is obvious that the choice of initial guess affects the convergence of the algorithm. Theoretically the method will always converge to the right eigenpair as long as the starting vector is close enough to the solution. In CI problems however, if the matrix contains some hidden symmetry and the starting vector is orthogonal to the desired root, the algorithm will fail. In practice sufficiently tight convergence thresholds are not always used for CI problems [25]. For the CI problem a common method of generating an initial guess is diagonalizing a small sub-matrix of the Hamiltonian [25]. Alternatively unit vectors $\mathbf{e}_i$ can be used and the simplest way of choosing them is to look at the size of the diagonal elements of the matrix. The smallest ones will be the ones of interest for the ground state. Using unit vectors means we are less likely to run into problems with hidden symmetry. This will not guarantee convergence to the lowest solutions, and a more practical approach is to get the algorithm to find more roots than are needed. They can be found with a soft convergence threshold of 5×10^{-3} say. (This is compared with the usual threshold of 10^{-4}.) The unwanted higher roots are discarded and the remaining roots are fully converged. It can be said for CI problems sufficiently good initial guesses are often available, which helps explain why Davidson's method works so well for these problems.

Finally we mention how Davidson's method can be used to solve the generalized eigenvalue problem $H\mathbf{c}=E\,S\mathbf{c}$. All that is needed is a new scalar product [5]. In practice the convergence is not nearly as good on these problems.

6.4 Some Recent Modifications Used For CI

We now discuss some modifications of Davidson's method that have been used for solving the CI eigenvalue problem. These modifications basically involve making some sort of approximation to reduce the storage requirements. This means that the methods are designed with large-scale CI problems in mind. The need for a space saving method is pointed out by Shepard in [40].

In this section we will continue to use the notation of the last couple of sections. However here the matrix A will be the Hamiltonian and the eigenvalues will represent the energy states. This means the matrix that is (partially) diagonalized is the representation of the Hamiltonian in a subspace. We are mainly interested in calculating the ground state so we want the smallest eigenvalue and its eigenvector.

We start by looking at van Lenthe and Pulay's modification [44], which is based on the conjugate gradient method. The conjugate gradient method is used for solving linear systems of equations not eigenvalue problems. It is applicable here because the approximation to the eigenvalue is assumed to be constant. We then briefly consider calculating higher eigenvalues and point out that the modification can be used for this too. Finally we look at some other alternatives.

6.4.1 The Modification of van Lenthe and Pulay

The modification we are about to look at was due to van Lenthe and Pulay in 1990 [44]. It only needs five to seven vectors stored. The exact number depends on how it is implemented. The fact that it only needs the storage of a few vectors means it can be used for very large problems, where disk storage is the limiting factor. It does not have to be restarted like Davidson's method does on large problems. The method as it was first stated was applicable only for the ground state. However it can be applied in the optimization of any excited state if it is started with an appropriate vector [3, page 24].

Before we look at the modification we note a couple of things about Davidson's original method. Firstly, the expansion vectors are orthogonalized. This is done to improve numerical stability and does not have to be done. Secondly for a *single* eigenvector Davidson's method requires the storage of $2j$ vectors, where j is the dimension of the expansion subspace at the current iteration. This is j approximate eigenvectors and j residues which require storing $A\mathbf{y}$ or $(A - \theta I)\mathbf{y}$. When this number gets too big the procedure is restarted. This can slow down convergence as each vector in the whole P_j space will contribute to the "best" vector at each iteration.

The following is an important idea behind the modification. In the iterative solution of the eigenvalue problem, convergence of the energy is quadratic and is therefore much faster than the convergence of the eigenvectors. So most of the iterations are to get the eigenvectors accurately. The modification assumes that the eigenvalue (and hence energy) is essentially constant, which means the eigenvalue equation is like a linear system.

Van Lenthe and Pulay use the approximation that the CI problem is a quadratic functional. Hence the applicability of the conjugate gradient method. With this restriction only *three* vectors are needed in the subspace [47] and this comes from conjugate gradient method theory [47]. Van Lenthe and Pulay show that as long as the eigenvalue has converged enough so that it is constant Davidson's method is equivalent to the conjugate gradient method [44]. Convergence in the conjugate gradient method is guaranteed only for positive definite systems, and in the CI context this is when we are looking for the lowest eigenpair.

At each step of Davidson's method the matrix Q_j grows by one column which, as we have already pointed out, can cause storage problems. The matrix Q_j crops up in the Rayleigh-Ritz procedure and it contains an orthonormal basis for the space span P_j. Now we look at which three vectors get used in van Lenthe and Pulay's algorithm [44]. At step j we might expect the most recent vectors to be used. In fact $\mathbf{q}_{j-1}$ and $\mathbf{q}_j$ are replaced by the Ritz vectors of the desired eigenpair, $\mathbf{y}_{j-1}$ and $\mathbf{y}_j$ respectively. The new vector $\mathbf{q}_{j+1}$ is evaluated in the same way as in Davidson's method. No orthogonalization is used in the whole iterative process, and this means we get a generalized eigenvalue equation like we did on page 116. However as it is only a 3×3 one it is not "hard" to solve. Bofill and Anglada [3] look at justifying this modification from the Lanczos algorithm. We will not go into any more detail here because it cannot be done properly without going into conjugate gradient theory.

We note that the van Lenthe and Pulay method should converge well if the approximation to the desired eigenvalue does not change much in the whole iterative process. A numerical analysis of this simplification has been given [25] and it is shown to be more effective than the original algorithm.

Murray, Racine and Davidson [25] suggest a modification slightly different to that of van Lenthe and Pulay. Truncation will still inhibit convergence in so far as the eigenvalue is not constant during the iterative process. They suggest that as much disk space as is available should be used provided that the I/O of the vectors does not start to become competitive with the time taken accessing or calculating the matrix.

6.4.2 Calculating Higher Eigenpairs

It is straightforward to extend Davidson's method to calculate higher energy states. The higher eigenpairs in the subspace are approximations to the higher eigenpairs of the full matrix. If we want the kth eigenpair of the full matrix then we can look at successive improvements of the Ritz vector from the kth eigenvector of the subspace matrix. This approach is guaranteed to work when the estimates of the $k-1$ lower eigenvalues produced by the subspace matrix diagonalization are smaller than the kth exact eigenvalue. This means that reasonable approximations to the lower eigenvectors have to be contained in the subspace. Suppose we want all the eigenvalues up to a certain number. As one of the major bottlenecks is reading from disk or recalculating the matrix elements it makes sense to add more than one vector at a time to the P space. This is the **block Davidson method** and it is similar to the block Lanczos method, which is described in [11, page 485–487]. For each vector added two vectors have to be stored in main memory. So not all the vectors can be considered at every iteration.

Murray, Racine and Davidson [25] suggest adding as many vectors as memory constraints allow and cycling through the unconverged roots in turn until all the desired eigenvectors have been converged. The procedure offers considerable savings in CPU time over the "one root at a time" approach. It is shown in [25] that the truncation procedure of van Lenthe and Pulay is effective when applied to excited states calculations when it is used with the block Davidson method.

6.4.3 Some Other Modifications

The solution of the CI equations is obtained by solving the problem in a given subspace. This subspace is usually increased in dimension at each iteration. The difference between between methods is the way the new vectors are generated. In this chapter we have only looked properly at two different ways of generating the new vectors, that in the Lanczos algorithm and Davidson's method. The DIIS algorithm of the last chapter is also a subspace method that is related to what we are talking about here, but only in a general way as it does not seem to get used for CI problems. We also looked at a subspace method in the parallel direct-SCF procedure. We now mention another two different ways of generating new vectors. The first is very different from that used in Davidson's method, and the second is just a different preconditioner from Davidson's method. Both can be used with van Lenthe and Pulay's truncation idea.

In Davidson's method one of the key steps is the minimization of the Rayleigh quotient in the orthogonal subspace generated in the iterative procedure. Another expansion is possible based on the minimization of the least squares error in the residual vector, and this is discussed by Murray, Racine and Davidson in the 1992 article [25]. It uses extrapolation. An attractive feature of it is that it concentrates on convergence of the vector rather than the eigenvalue. So the method could be used to complement Davidson's method. They show that the Davidson algorithm for large matrix diagonalization can be made to converge more quickly by periodic use of this least squares extrapolation.

In a 1994 paper [3] Bofill and Anglada look at using a preconditioning different from that used in Davidson's method. It is based on the fact that the eigenvalue problem can be seen as a stationary condition on the Lagrangian function

$$\mathcal{L}(\mathbf{y}_j, \lambda_j) = \mathbf{y}_j^T H \mathbf{y}_j - \lambda_j \left[\mathbf{y}_j^T \mathbf{y}_j - 1\right]$$

It was shown to substantially reduce the number of iterations on some problems [3].

We now very briefly mention an alternative method [40] that saves space for large-scale eigenvector problems. It is a data compression method and was due to Shepard in 1990. It has nothing to do with the modifications we have just looked at. It is based on reducing the accuracy of the eigenvectors in the later stages of the iterations. Basically the data is compressed by lowering the precision of the stored expansion vectors. If the expansion vectors can be stored in the main memory of the computer then the I/O requirements can be greatly reduced. On a particular type of CI problem it was predicted that if the size was 10^8, truncation from 64-bit to 15-bit floating point representation would give the asymptotic full-precision convergence rate [40]. This means that in the first "few" iteration the full-precision (64-bit) algorithm converges more quickly, but as the iterations proceed the rates become approximately the same. However in the full-precision case we get closer to the solution before the rate slows down. The I/O requirements have obviously been reduced significantly but the number of iterations taken to reach convergence is increased. It seems worthwhile overall though. For further details of this method the reader is referred to [40].

Chapter 7

Summary

In this thesis we have mainly focused on two eigenvalue problems from quantum chemistry. We now review what we have been looking at.

The first of these eigenvalue problems was the Roothaan equations. These are used for calculating the Hartree-Fock wave function in the restricted closed-shell case. We went through the derivation of these equations by starting with Schrödinger equation that we want to solve. They come from minimizing the energy of the system. We are mainly interested in calculating the energy of the ground state. The equations are a generalized eigenvalue problem of the form

$$F\,C = S\,C\,\varepsilon$$

where F is the Fock matrix, C is the expansion coefficient matrix, S is the basis function overlap matrix and ε is a diagonal eigenvalue matrix. The eigenvalues represent ionization potentials and electron affinities according to Koopmans' theorem. We looked at two ways of solving these equations. The first was Roothaan's self-consistent field procedure (1951). It is an iterative method that can be expressed as

$$F[C^{(i)}]\,C^{(i+1)} = S\,C^{(i+1)}\,\varepsilon^{(i+1)}$$

The focus of the scheme is getting a self-consistent Fock matrix. In the equation above this means $C^{(i)} = C^{(i+1)}$. There is a diagonalization step involved. This simple iterative scheme has three main problems. The first is that it does not always converge. The level-shifting method (1973) is used to ensure convergence of the SCF procedure. Another problem is slow (linear) convergence and direct inversion in the iterative subspace (1982) is used to speed up SCF convergence. The final problem is storage. For problems of a decent size there are too many two-electron integrals to be stored in the CPU of a computer. This has resulted in the direct-SCF procedure which is so called because the matrix elements are calculated *directly* from the basis functions. The second method we looked at was a direct-SCF method that is second-order and implemented in parallel (1993). The focus here is on minimizing the energy.

The Hartree-Fock approximation to a great extent neglects the fact that the motion of electrons is correlated and electrons tend to avoid each other. We looked at going beyond this approximation and the idea of configuration interaction. This used a linear combination of configurations to improve the wave function. The particular method that we focused on most closely was the configuration interaction method. Here the secular equations

$$H\mathbf{c} = E\,S\mathbf{c}$$

are solved. The matrix H is a Hamiltonian and gives the interactions between the configurations we are considering. The matrix S is the configuration overlap matrix. The eigenvector $\mathbf{c}$ contains the configuration expansion coefficients and the eigenvalue E is the energy. With this problem it is relatively easy to get a transformation to get S equal to the identity. This gives us the second eigenvalue problem we looked at

$$H\mathbf{c} = E\,\mathbf{c}$$

There is a direct version of this method (1972) as well and the elements of H are constructed directly from the molecular orbitals that make up the configurations. Davidson's (original) method (1975) is the most common method that gets used to solve this eigenvalue problem. It works well because H has a dominant but non-constant diagonal, and only 1–5% of its elements are randomly non-zero. Only one or very few of the lowest eigenvalues and corresponding eigenvectors tend to be needed. The generalization (1986) and modification of Davidson's method were discussed. In particular a modification due to van Lenthe and Pulay (1990) which is used to cut down storage requirements for the CI eigenproblem was looked at.

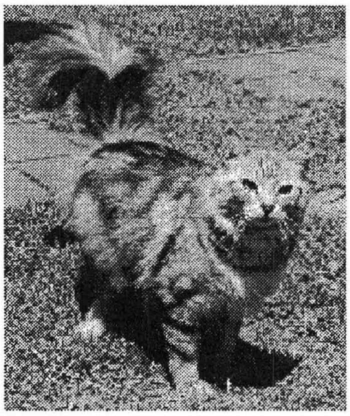
Acknowledgements

A huge thank you to my supervisor John Hannah for all his support, encouragement and proof-reading.

I am grateful to Dr. Christian Bischof of Argonne National Laboratory for sending me a copy of [2].

Thank you to Professor George Schatz of Northwestern University for being so helpful with references that were useful in getting started.

I am indebted to my Warble for his enthusiasm, LaTeX help and millions of other things. ☺



Bibliography

[1] Bacskay, G. B. [1981] A quadratically convergent Hartree-Fock (QC-SCF) method. Application to closed shell systems *Chem. Phys.* **61**, 385–404

[2] Bischof, C. H., Shepard, R. L., Huss-Lederman, S. (Editors) [1996] *Workshop Report on Large-Scale Matrix Diagonalization Methods in Chemistry Theory Institute* Tech. Rep. MCS-TM-219 (Argonne National Laboratory)

[3] Bofill, J. M., Anglada, J. M. [1994] Some remarks on the use of the three-term recurrence method in the configuration interaction eigenvalue problem *Chem. Phys.* **183**, 19–26

[4] Császár, P., Pulay, P. [1984] Geometry optimization by direct inversion in the iterative subspace *J. Mol. Struct.* **114**, 31–34

[5] Davidson, E. R. [1989] Super-matrix methods *Comp. Phys. Comm.* **53**, 49–60

[6] Davidson, E. R. [1975] The Iterative Calculation of a Few of the lowest Eigenvalues and Corresponding Eigenvectors of Large Real-Symmetric Matrices *J. Comput. Phys.* **17**, 87–94.

[7] Diercksen, G. H. F., Wilson, S. [1983] *Methods in Computational Molecular Physics* (D. Reidel Publishing Company, Dordrecht Holland)

[8] Fock, V. [1930] Näherungsmethode zur Lösung des quantenmechanischen Mehrkörperproblems *Z. Phys.* **61**, 126–148

[9] Gilbert, G., Smith, S. C. [1990] *Theory of Unimolecular and Recombination Reactions* (Blackwell Scientific Publications, Oxford)

[10] Grein, F., Chang, T. C. [1971] Multiconfiguration wavefunctions obtained by application of the generalized Brillouin theorem *Chem. Phys. Lett.* **12**, 44–48

[11] Golub, G. H., Van Loan, C. F. [1996] *Matrix Computations 3rd Edition* (The John Hopkins University Press, Baltimore)

[12] Hall, G. G. [1951] The molecular orbital theory of chemical valency VIII. A method of calculating ionization potentials *Proc. Roy. Soc. (London)* **A205**, 541–552

[13] Hamilton, T. P., Pulay, P. [1986] Direct inversion in the iterative space (DIIS) optimization of open-shell, excited-state, and small multiconfiguration SCF wave functions *J. Chem. Phys.* **84**, 5728–5734

[14] Harrison, R. J., Shepard, R. [1994] *ab initio* electronic structure on parallel computers *Annu. Rev. Phys. Chem.* **45**, 623–658

[15] Hartree, D.R. [1928] The Wave Mechanics of an Atom with a non-Coulomb Central Field. Parts I, II, III *Proc. Camb. Phil. Soc.* **24**, 89–110, 111–132, 426–437

[16] Hirst, D. M. [1990] *A Computational Approach to Chemistry* (Blackwell Scientific Publications, Oxford)

[17] Huss-Lederman, S. [1996] Large-Scale Matrix Diagonalization In Computational Chemistry *SIAM News* **29**, 14,17

[18] Koopmans, T. A. [1933] Über die zuordnung von Wellenfunktionen und eigenwerten zu den einzelnen elektronen eines atoms *Physica* **1**, 104–113

[19] Lawley, K. P. (Editor) [1987] *Ab Initio Methods in Quantum Chemistry Part* II (John Wiley & Sons, Chichester)

[20] Levine, I. N. [1983] *Quantum Chemistry 3rd Edition* (Allyn and Bacon Inc., Newton Massachusetts)

[21] Levy, B., Berthier, G. [1968] Generalized Brillouin Theorem for Multiconfigurational SCF Theories *Int. J. Quant. Chem.* **2**, 307–319

[22] Lipkowitz, K. B., Boyd, D. B. (Editors) [1996] *Reviews in Computational Chemistry* (VCH Publishers, New York)

[23] Mattson, T. G. (Editor) [1995] *Parallel Computing in Computational Chemistry* (American Chemical Society, Washington)

[24] Morgan, R. B., Scott, D. S. [1986] Generalization of Davidson's method for computing eigenvalues of sparse symmetric matrices *SIAM J. Sci. Stat. Comput.* **7**, 817–825

[25] Murray, C. W., Racine, S. C., Davidson, E. R. [1992] Improved Algorithms for the Lowest Few Eigenvalues and Associated Eigenvectors of Large Matrices *J. Comput. Phys.* **103**, 382–389

[26] Parlett, B. N. [1980] *The Symmetric Eigenvalue Problem* (Prentice-Hall, Englewood Cliffs)

[27] Pulay, P. [1982] Improved SCF Convergence Acceleration *J. Comput. Chem.* **3**, 556–560

[28] Pulay, P. [1980] Convergence acceleration of iterative sequences. The case of SCF iteration *Chem. Phys. Lett.* **73**, 393–398

[29] Roos, B. O., Taylor, P. R., Siegbah, P. E. M. [1980] A CASSCF method using a density matrix formulated super-CI approach *Chem. Phys.* **48**, 157–173

[30] Roothaan, C. C. J. [1960] Self-consistent field theory for open shells of electronic systems *Rev. Mod. Phys.* **32**, 179–185

[31] Roothaan, C. C. J. [1951] New developments in molecular orbital theory *Rev. Mod. Phys.* **23**, 69–89

[32] Saad, Y. [1992] *Numerical methods for large eigenvalue problems: theory and algorithms* (Manchester University Press, Manchester UK)

[33] Saunders, V. R., van Lenthe, J. H. [1983] The direct CI method. A detailed analysis *Mol. Phys.* **48**, 923–954

[34] Saunders, V. R., Hillier, I. H. [1973] A "Level-Shifting" Method for Converging Closed Shell Hartree-Fock Wave Functions *Int. J. Quant. Chem.* **7**, 699–705

[35] Schaefer, H. F. (Editor) [1977] *Modern theoretical chemistry; volume 3: Methods of electronic structure theory* (Plenum Press, New York)

[36] Sellers, H. [1993] The C^2 – DIIS Convergence Acceleration Algorithm *Int. J. Quant. Chem.* **45**, 31–41

[37] Sellers, H. [1991] ADEM-DIOS: an SCF convergence algorithm for difficult cases *Chem. Phys. Lett.* **180**, 461–465

[38] Shepard, R. [1993] Elimination of the diagonalization bottleneck in parallel Direct-SCF methods *Theor. Chim. Acta* **84**, 343–351

[39] Shepard, R. [1992] On the global convergence of MCSCF wave function optimization: The method of trigonometric interpolation *Theor. Chim. Acta* **84**, 55–83

[40] Shepard, R. [1990] A Data Compression Method Applicable to First-Order Convergent Iterative Procedures *J. Comput. Chem.* **11**, 45–57

[41] Stewart, J. J. P., Császár, P., Pulay, P. [1982] Fast Semiempirical Calculations *J. Comput. Chem.* **3**, 227–228

[42] Szabo, A., Ostlund, N. S. [1989] *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory, First Edition Revised* (McGraw-Hill Inc., New York)

[43] Trefethen, L. N., Bau III, D. [1997] *Numerical Linear Algebra* (SIAM Publications, Philadelphia)

[44] van Lenthe, J. H., Pulay, P. [1990] A Space-Saving Modification of Davidson's Eigenvector Algorithm *J. Comput. Chem.* **11**, 1164–1168

[45] Whitten, K. W., Gailey, K. D., and Davis, R., E. [1988] *General Chemistry with Qualitative Analysis 3rd Edition* (Saunders College Publishing, Philadelphia)

[46] Wolinski, K., Haacke, R., Hinton, J. F., Pulay, P. [1997] Methods for Parallel Computation of SCF NMR Chemical Shifts by GIAO Method: Efficient Integral Calculation, Multi-Fock Algorithm, and Pseudodiagonalization *J. Comput. Chem.* **18**, 816–825

[47] Wormer, P. E. S., Visser, F., Paldus, J. [1982] Conjugate Gradient Method for the solution of Linear Equations: Application to Molecular Electronic Structure Calculations *J. Comput. Phys.* **48**, 23–44

[48] Wu, X. T., Hayes, E. F. [1997] Algorithms for Obtaining Cumulative Reaction Probabilities for Chemical Reactions *J. Comput. Phys.* **130**, 136–147

[49] Wyatt, R. E., Iung, C., Leforestier, C. [1995] Toward *ab Initio* Intramolecular Dynamics *Acc. Chem. Res.* **28**, 423–429

[50] Yu, H. G., Smith, S. C. [1997] Restarted Krylov-space spectral filtering *J. Chem. Soc. Faraday Trans.* **93**, 861–869